UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

WINDOWS VISTA MEDIA CENTER CONTROLLED BY SPEECH

Mário Miguel Lucas Pires Vaz Henriques

Mestrado em Engenharia Informática

2007

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

WINDOWS VISTA MEDIA CENTER CONTROLLED BY SPEECH

Mário Miguel Lucas Pires Vaz Henriques

Projecto orientado pelo Prof. Dr. Carlos Teixeira
e co-orientado pelo Prof. Dr. Miguel Sales Dias

Mestrado em Engenharia Informática

2007

Acknowledgments

I would like to thank…

To Prof. Doutor Miguel Sales Dias for his orientation and supervision and for allowing my presence in the MLDC team;

To Prof. Doutor Carlos Teixeira for his orientation and supervision during my project;

To Eng. Pedro Silva for his amazing friendship, constant support and help;

To Eng. António Calado by his constant help and support;

To Sandra da Costa Neto for all the support and help executing the usability tests;

To all other colleagues from MLDC;

To all Microsoft colleagues that participated in the usability tests.

Contents

Index of Figures

Acronyms

AEC *– Acoustic Echo Cancellation*
API *– Application Programming Interface*
ASR *– Automatic Speech Recognition*
EMEA *– Europe, Middle-East and Africa*
ENG *- English*
EVA *– Easy Voice Automation*
HCI *- Human-Computer Interface*
HMM *– Hidden Markov Model*
LAN *– Local Area Network*
MC *– Microsoft Media Center*
MLDC *– Microsoft Language Development Center*
PTG *– European Portuguese*
SAPI *– Microsoft Speech API*
SDK *– Software Development Kit*
SR *– Speech Recognition*
TTS *– Text-to-Speech*
WMP *– Windows Media Player*

1. Introduction

1.1 Introduction

Automatic Speech recognition (ASR) and synthesis technologies are already present in the daily life of an increased number of people. Due to the potential of speech, as a natural Human-Computer Interface (HCI) modality, a large number of applications are adopting it for their main control interface. Commercial, educational and medical [SPEECHAPP] applications are taking benefits of speech, bringing not only a more sophisticated interface for command and control and sometimes dictation, but also broadening its usage for all kinds of users, including children, elder and visual impaired persons.

ASR technology is nowadays helpful for many professional contexts, home entertainment and in small daily tasks that contribute for the general well-being. Applications such as Voice Command [VOICECOMMAND] for Windows Mobile [WMOBILE], Windows Vista Speech Recognition [SRAPI] for operating system full control, or telephony server-based applications, are some examples of products now available using speech technology, that are useful to simplify people's life.
However, as for the home "living-room" scenario, whereas a user accesses the home audio-visual devices via natural HCI modalities such as speech, there are very few applications that use speech recognition and synthesis, described in the literature. One such project, referred to as "Dee Jay" [DEEJAY], is a voice-controlled juke box for Windows Vista: you tell it what song you want to hear; the corresponding song entry is found in your library and the Windows Media Player [WMP11] will play it. It is also capable of placing searches with different criteria: by artist or genre, play albums or collections, and perform a number of other commands. This application only enables listening to music using Media Player.
Our line of work, extends this prior knowledge, by supporting fully the "living-room" scenario (section 2.3 The Usage Scenario), were different kind of media such as TV, radio, still and moving pictures, music and internet access are available.

Windows Vista Media Center (or MC in short) provides all-in-one solution for this kind of scenario. This platform, available in Microsoft Windows Vista Ultimate or Home Premium, gathers all digital entertainment (live and recorded TV, music, movies and pictures) in one device with a simple user-friendly interface totally controlled by a traditional remote control.
Introducing Automatic Speech Recognition technology in this platform will make possible for Media Center users to have a total control of the system, using for that purpose only their pronounced speech, giving a secondary role to the remote control interface.
Once finished, this speech interface for Media Center control would be available for Portuguese and English users.

This thesis summarizes all the work developed in the "Microsoft Media Center Controlled by Speech" project by a student of Faculdade de Ciências da Universidade de Lisboa, on the curricular period of his Masters in Computing Engineering (Mestrado em Engenharia Informática), developed in Microsoft premises in Portugal, Tagus Park, Porto Salvo, at the Microsoft Language Development Center.

Before starting describing the project itself, it is useful for the reader to be familiarized with some fundamental concepts, including Automatic Speech Recognition and Microsoft Windows Vista Media Center platform. These general ideas will be presented in the following section.

1.2 Automatic Speech Recognition

Automatic Speech recognition [ASR], from a simple view, is the process of converting a speech signal pronounced by a human user, to a symbolic representation like text or even actions, through computer algorithms, without human intervention.
Most of speech recognition systems are generally based on Hidden Markov Models (HMMs). The HMM [HMM] is a statistical model with a finite set of states, each one associated with a probability distribution. Transitions between states are governed by a set of probabilities called transition probabilities.

Speech recognition systems, depending on several different factors, could have a very high performance in controlled conditions. These factors include the amount of noise in the environment, the speaking rate of the speaker or the context (or grammar) being used in recognition.
The main goal of Automatic Speech Recognition technology is to help not only users with accessibility problems, but also to support users on their daily tasks, with a natural Human-Computer Interface.
For this project development we have adopted the existent Microsoft Speech Technology [MSPEECH07], including the new Portuguese Speech Recognition Engine for Client ("Speech Recognition Sample Engine for Portuguese (Portugal) [1], recently developed by the Speech Portuguese team at Microsoft [MLDC].

The ASR technology has experienced great progress over the past years. Since 1936, when the research of this technology began, several breakthroughs have contributed for his progress. One of the most important occurred in the early 70's, with the Hidden Markov Models approach to speech recognition. The responsible researcher was Lenny Baum of Princeton University, shared with several ARPA (Advanced Research Projects Agency) contractors.
In 1971, the first computer system that could understand continuous speech was developed by DARPA (Defense Advanced Research Projects Agency in USA). Since then, the speech recognition technology has been progressing following the computer science and computing architecture progress (speed and power).
Nowadays, this technology has already become a strong competitor of the traditional HCI's (Human Computer Interfaces), like remote control, mouse, keyboard and touch screens, and is helpful for many professional contexts, home entertainment as well as in the small daily tasks that contribute for the general well-being. There are several different applications that take use of this HCI modality, like Voice Command, Windows Mobile or even telephony server-based applications.

1 - http://www.microsoft.com/portugal/mldc/betaprograms/winclientdesktop.mspx

1.3 Microsoft Windows Vista Media Center

ASR will be applied to Microsoft Windows Vista Media Center [VMC07] (or Media Center in short), which is a Microsoft platform for home digital entertainment, with an easy-to-use interface appropriate for all family members.

Media Center (Figure 1) is designed for a wide range of displays (TV, Plasma, LCD, HD-TV, etc) and input methods, such as a remote control or mouse and keyboard, touch screen, or even a Tablet PC.
With this platform it is possible to view photos in a slide-show, browse and play a music collection, easily play movies and DVD's, watch and record live TV, listen to radio, download music and videos, and so on. The main use case for this platform is the "living room" scenario, where users can comfortably take control of all these features.

Media Center is available since the release of the previous Microsoft Windows XP (25th October 2001) and it was improved since (with more functionalities; improved user interface; more extensibilities; support for digital cable service), in the new Microsoft Operating System Windows Vista release (available in Ultimate and Home-Premium editions).

From a global view, the main propose of this project is to develop a general speech interface, in both Portuguese and English languages, to control Microsoft Media Center, using current Microsoft Speech Recognition technology.

By introducing Speech technology in this kind of platform, we are envisaging a more natural and simple control of its functionalities. Spoken commands like "Play Jazz", "Watch channel 4" or maybe "Show 2005 vacation pictures" introduce a natural and simple way to interact with the system and control its functionalities, giving a secondary role to the remote control UI. This kind of interaction increases also the accessibility for users groups that are often forgotten, such as the elderly and young children.

Figure 1 – Media Center logo

1.4 Host institution

This thesis was developed in Microsoft Portugal premises at Tagus Park, in the MLDC (Microsoft Language Development Center) Development Center.
Microsoft [MSFT07] is nowadays the largest reference in the world regarding technology and computer science.

The company effort is to provide users of its platforms and systems, tools to aid the resolution of their daily problems, not only to the professional level but also to the personal level.

During the time I have been working at Microsoft, I have perceived that one of the secrets of the company for success relates to its internal policies that grant the employees and collaborators a perfect working environment, resulting in a willingness of these to help their peers in their daily work, thus providing the conditions to satisfy their customers.

Within Microsoft Portugal, I joined the MLDC development center (Microsoft Language Development Center - Portugal) [MLDC06]. The best way we found to describe this Center (Figure 2) is by the words of its main responsible, Prof. Doutor Miguel Sales Dias.

Figure 2 – MLDC logo

"MLDC is the first Microsoft Development Center outside of Redmond dedicated to key Speech and Natural Language developments, and is a clear demonstration of Microsoft efforts of stimulating a strong software industry in EMEA. To be successful, MLDC must have close relationships with academia, R&D laboratories, companies, government and European institutions. I will continue fostering and building these relationships in order to create more opportunities for language research and development here in Portugal.

Miguel Sales Dias, Director of the Microsoft Language Development Center"

1.5 Thesis problems identification

This thesis's main goal was to develop a new functionality to the Media Center system - the possibility for users to have full system control using only their pronounced speech in Portuguese or in English. The development of this new functionality generated some complex problems that had to be solved during this project execution.

Regarding the Media Center platform and speech recognition technology for a living room scenario, two major problems were identified.

The first one is related with the automatic language detection, where any user has the possibility to use Portuguese or English language for system control. In this case, the problem is related to the language detection, in real time, that it being used by the user in a specific moment.

The second and more complex problem is the cancellation of the audio source, for robust speech recognition in arbitrary acoustic scenarios (such as the living room scenario), due to the audio output from music or film.

1.6 Solution approaches

For the two major problems identified we needed to find some approaches.

1.6.1 Automatic language detection

For the automatic language detection, between Portuguese and English, one of the possible approaches could be the simultaneous functioning of the two recognition engines. By following this approach, we simply choose to put the recognition engines working at the same time (concurrently), using their own grammars. To prevent some misunderstood commands (Portuguese to English or vice-versa) we simple avoid similar commands in the grammars (Annex 2 and 3). For example, to enter in the videos menu we choose "video library" command to the English grammar and "videos" for the Portuguese grammar.

1.6.2 Acoustic Echo Cancelation

To have perceived speech commands, the system needs to have the capacity "to understand" when the user is sending an order, even if the music or TV sound is too high. For this reason, we started the enhancement of the living room scenario by coping with Acoustic Echo Cancellation and Noise Reduction, trough Automatic Gain Control, which are already incorporated in Microsoft Microphone Array technology (which includes a set of closely positioned microphones) [MicArray06]. The Microphone Array is already supported in Windows Vista, but not yet in Speech Recognition. Our line of work, in a first phase was the integration of MicArray in Speech Recognition Engine, and in a second phase the integration of the latter in Media Center. With the integration of this technology in the Media Center Controlled by Speech, the system should be able to capture user's speech commands with Acoustic Echo Cancellation (AEC), for robust speech recognition of a speaker in the presence of ambient noise and/or music. This technological approach is described, in detail, later in chapter 3.3.2 Microphone Array.

1.7 Main thesis goals

The main goal of this thesis was the development of a command and control speech interface, in both European Portuguese and English languages, for the Microsoft Media Center platform. To be successful it was necessary to solve the problems identified in the previously chapter: the Automatic language detection and acoustic echo cancelation.
So, at the end of this thesis, the following goals would have to be achieved:

1. Automatic language detection, between English and Portuguese,

2. Robust speech commands to control the system, in English and Portuguese, supporting complex "speech macros" commands like: "Turn on TV", "change to channel 2", "play jazz music", "show 2006 vacations photos", "play Sting", and so on,

3. Performing usability tests to determine the usability of the system;

4. Acoustic Echo cancellation support using Microphone Array technology for robust speech recognition. Integration in the Speech Recognition Engine (in a first phase) and integration in Media Center (in a second phase/ future work).

When all these goals were accomplished, the system will be able to function in usage scenarios such as Media Center controlling the TV, playing some music (Figure 3) or a film.

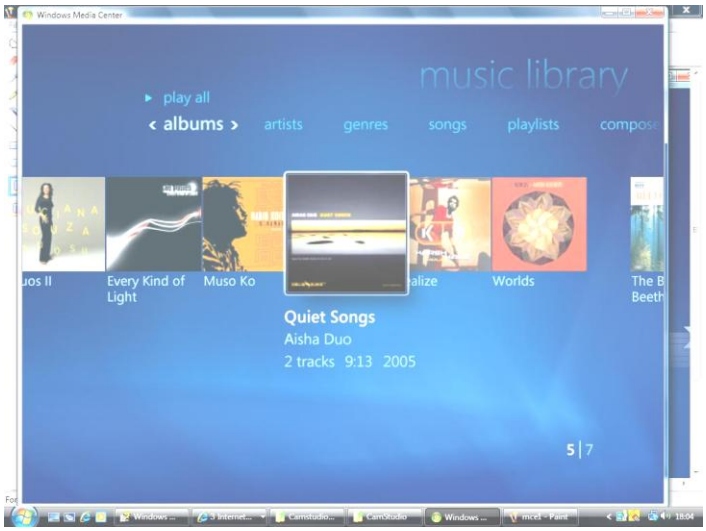


Figure 3 – Media Center music menu

1.8 Chapter Organization

This thesis is organized as follows:

Chapter 1, Introduction

This is the current chapter in which we introduce the thesis objective, the state of the art and the main technologies involved for the project development, the Automatic Speech Recognition and the Windows Vista Media Center.
In this chapter we also identified and analyzed the major problems concerning the speech interface development; we present solution approaches for each one of the identified issues and finally we show the list of the most important tasks that were accomplished during this thesis development.

Chapter 2: User and System requirements

This chapter describes all the user and system requirements, specifying hardware and application detailed requirements. The usage scenario for this kind of system, using a speech interface, is also presented in this chapter.

Chapter 3: System Architecture and development

In this chapter is presented the complete application architecture, including a detailed description for each module, as well as all the application development description, covering all the requirements identified in chapter 2.

Chapter 4: Usability Evaluation

In this chapter we present all the information related to the usability tests conducted on the system using the new speech interface for Media Center, describing the methodology, evaluation results, subject's comments and feedback, analysis and conclusions.

Chapter 5: Conclusions and Future Work

In the last chapter is made the thesis conclusions and final remarks. Possible future work directions and lines of research are also discussed.

1.9 Conclusion

In this initial chapter we have introduced the reader to the thesis background, namely the Automatic Speech Recognition technology and the Windows Vista Media Center platform. The main objective of this thesis was also presented, as well as other similar applications that take benefit of using speech recognition on their main interface.

The development of a speech interface for this kind of systems generates some complex issues that needed to be solved. The problems were the automatic language detection and the Acoustic Echo Cancellation for robust speech recognition due to the audio output from the system. In this chapter we have identified the main thesis goals that needed to be accomplished to successfully complete this work.

For the automatic language detection the approach was the simultaneous functioning of two different recognition engines. By following this approach we had to construct two different grammars where there were no similar commands, avoiding misunderstood commands.

For the Acoustic Echo Cancellation problem the approach was using Microphone Array. The work was the integration in the Speech Recognition Engine (in a first phase) and the integration in Media Center (in a second phase/ future work).

2. User and System Requirements

2.1 Introduction

Microsoft Media Center is a platform for family digital entertainment at home. The main objective of this system is to concentrate in one single technology and device the capability to explore and view all audio-visual media. This platform has a simple interface, as showed in Figure 4, which is currently managed by a remote control. The large number of menus and options offered by the system increases the complexity for the use of this traditional HCI (remote control, see Figure 5).

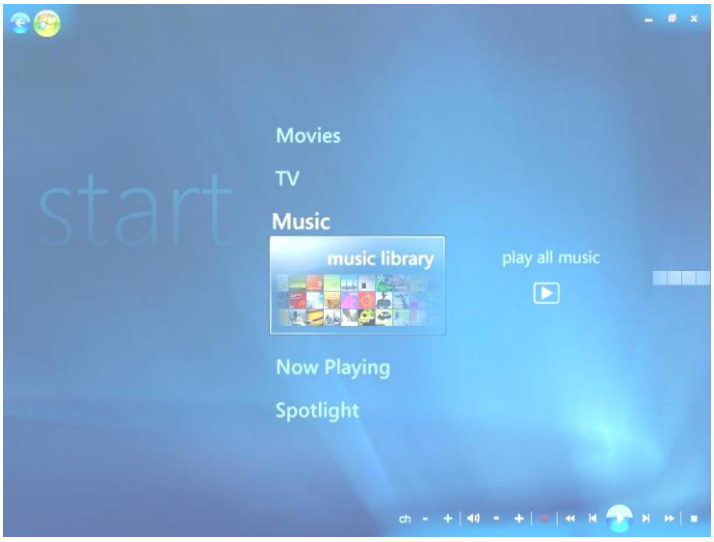


Figure 4 – Media Center simple User Interface

Figure 5 – MC remote control

In this chapter, we describe the full list of requirements captured for the system, specifying hardware, software and user requirements.

2.2 Overall user requirements

The main goal for the thesis was to allow Media Center full control using only user's pronounced speech, giving a secondary roll to the traditional HCI remote-control. The user should be able to control the main commands available, not only for menu navigation, but also to control all media available on the host machine. For this reason, commands like "play rock", "change to channel 4", "watch 2006 vacations photos", "play Sting", and so on, should be available to control the system.

Using the speech interface, it would be possible to control the system easily and more naturally than using the remote control. This fact increases the system accessibility, opening it for other kind of users, like children, elder and visual impaired persons. So, the target audience for this system (with speech interface) would be extended for all kind of Media Center users.

To improve the speech interface and increase the system accessibility, we also included speech macro-commands, unavailable in the remote control, where the user can perform complex commands using only 2/3 words, like "show vacation pictures" while user listen to some music.

In order to achieve this purpose, we have developed a speech interface, very simple to use, adapted to a large number of different users, with more than 300 different commands for just 3 major media areas – music, pictures and movies, and with specific media control commands.

2.3 The Usage Scenario

The usage scenario for Media Center is the living room, as depicted in Figure 6. This simple audio-visual system is connected to a TV for presentation of audio-visual content and to an audio distribution system (mono, stereo or surround), enabling users to watch and listen to their favorite movie, music, pictures and to access the Internet via a web browser.

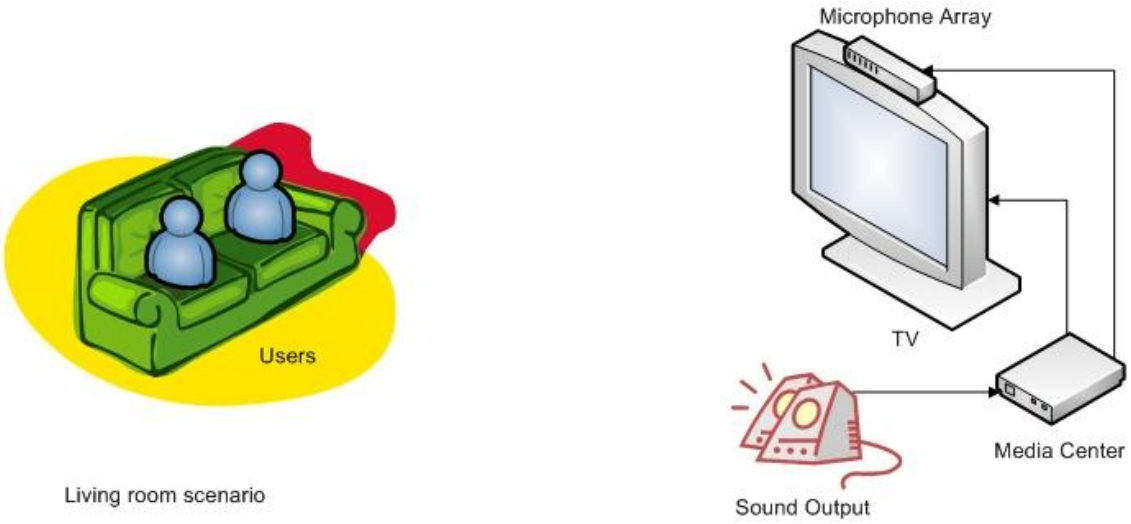


Figure 6 – The usage scenario

2.4 Detailed system requirements

2.4.1 Hardware Requirements

The main goal of Media Center platform was to provide support for any kind of media entertainment through a simple and small hardware component connected to a TV.

The Media Center provides support for various kinds of media entertainment through a simple and small hardware component connected to a CRT/LCD TV or monitor (Figure 8). This may be seen as a powerful PC (Figure 7) that must be capable of supporting the required types of digital media for entertainment purposes. For this reason, this hardware must fulfill the technical requirements showed in Table 1.

	Minimum Requirements	Recommended Requirements
Processor	800 MHz	1GHz 32-bit (x86) or 64-bit (x64)
System Memory	512 MB	1 GB
Hard Drive	20 GB	40 GB
Graphics	Support for SVGA	Support for DirectX
TV Tuner Card	Required	Required
Monitor	Any CRT or LCD monitor with S-Video, DVI or RGB-VGA connection	Any CRT or LCD monitor with S-Video, DVI or RGB-VGA connection
Microphone	Wireless headset or standard desktop microphone	Wireless headset, standard desktop microphone or Linear 4 Microphone Array
Operating System	Vista Ultimate or Home Premium	Vista Ultimate or Home Premium

Table 1 – Media Center PC technical requirement

Figure 7 - Media Center box

Figure 8 – High definition TV

For system control in a living room scenario, using speech, it is recommended a wireless headset microphone (Figure 10) or a standard desktop microphone (Figure 11). Using this kind of hardware users can easily control the system by sending speech orders directly to the system.

Later, with the Microphone Array (Figure 9) integration in Media Center, it will be possible users to speak freely to the system without using any headset microphone. This component will be connected to Media Center box and its function is to capture user's speech commands using Acoustic Echo Cancellation (AEC) for better recognition.

Figure 9 – Linear four Microphone Array

Figure 10 – Wireless headset microphone

Figure 11 – Standard Desktop Microphone

We have identified 3 major requirements at the hardware level, as showed in Table 2.

ID	DESCRIPTION
Media Center Box	The Media Center PC should be able to support various kinds of media entertainment through a simple and small hardware.
TV	The CRT/ LCD monitor should be able to show clearly all the system output to the users.
Microphone	The microphone should be able to receive user's orders and send them to the system. This microphone should present Noise Suppression capability for a better recognition.

Table 2 – Hardware Requirements

2.4.2 Application requirements

The speech interface for Media Center platform should allow users to control the system easily and more naturally than using the remote control and have total control of the latter with speech, in command and control mode, in European Portuguese, English or both languages in sequence. This interface should also allow the usage of Speech Macros, improving the usability of the system. These requirements were identified in the user's requirements and for these reasons we have identified 3 major requirements at the software level.

ID	TYPE	DESCRIPTION
SR01	Speech Recognition	The user should be able to interact with the system with simple speech commands like "play genre Jazz", "play all Sting music", "show channel 4" or even "show kids pictures". All these commands should be available in the grammars. Microsoft Speech Recognition technology in Portuguese and English should be used
MC01	Media Control	The user should be able to request all media available (music, pictures and movies) in the host machine. Media Player API should be used to retrieve all media information available in the host machine, and this information should be included in the grammars (see Table 5).
MCControl01	Media Center Control	All the Media Center main controls should be available to the user (see Table 4).

Table 3 – Overall Application Requirements

MEDIA TYPE	MAJOR AVAILABLE COMMANDS
Music	Play; Stop; Resume; Next Music; Previous Music; Forward; Rewind
Pictures	Show; Next Picture; Previous Picture
Videos	Play; Pause; Stop; Resume; Slower; Faster
MC control	Main Menu; Videos; Pictures; Music; Volume up; Volume Down; Mute;

Table 4 – Media Center controls

As identified in the application requirements, the system should allow user's to be able to perform the most common operations, like system and media control, in a simple and natural way, using for that only their pronounced speech. As we can see in Table 5, it should be possible more than one way to invoke a command, using for that several command synonymous. In annex 2 and 3, we show the product user-guide, where we present all the available commands.

ID	TYPE	DESCRIPTION	EXAMPLE
MUSIC01	MUSIC	Music selected starts playing: Tocar <*nome*> Ouvir <*nome*> Reproduzir <*nome*> <Nome do Media> <Tag do Media>	"Tocar estilo rock" "Ouvir artista Sting" "Sting" "Rock"
PICTURES01	PICTURES	Starts picture Slideshow: Mostrar <directório> Ver <directório> <directório>	"Mostrar fotos do Dinis" "Ver imagens do Afonso" "Férias de 2006"
VIDEOS01	VIDEOS	Video starts playing: Ver <nome> Mostrar <nome> <nome>	"Ver filme do verão" "Mostrar vídeo do urso" "Futebol"
MENU NAVIGATION01	MENU NAVIGATION	Media Center basic navigation: Menu principal Vídeos Música Imagens Mais programas Voltar	
MAIN COMMANDS01	MAIN COMMANDS	Always available to control Media Center: Mute Stop Volume Down Volume Up	

Table 5 - Detailed Application Requirements

2.5 Conclusion

In this chapter we have presented the user requirements as well as the hardware and the application requirements that were captured at the beginning of this work.
We have also presented the typical usage scenario for this kind of system using a speech interface.

3. System Architecture and development

The system uses Speech Recognition Technology with Portuguese/English acoustic models, developed at Microsoft, and also TTS (Text to Speech) technologies on those two languages (where the European Portuguese TTS is licensed from a third party Vendor).

SpeechFX (Speech library included on .Net Framework 3.0) [SFX07] has been used in the project development. The latter is a managed code API developed by Microsoft to allow the use of Speech Recognition and Speech Synthesis within Windows applications, using the .NET framework. EVA (Easy Voice Automation), a recent Microsoft technology (described latter in chapter 3.4.2), was also assessed for use in this project, but after some tests we have chosen SpeechFX.

3.1 Technologies used

For the Media Center speech interface development, a group of technologies were used. The principal component was the Media Center Add-In, where speech recognition and synthesis is made. The complete list of software, libraries and SDK´s used for the application development, is as follows:

- Microsoft Windows Vista Ultimate Operating System
- Microsoft Visual Studio 2005
- Windows Vista Media Center SDK
- Windows Media Player SDK
- Microsoft Speech FX library included on .Net Framework 3.0

In the next sub-sections we describe in more detail, the most relevant components used.

3.1.1 Vista Media Center SDK

The Media Center Add-In development requires the Media Center software development kit (SDK). This SDK is public [MCSDK07] and is designed to help developers to create applications and software components that take advantage of features provided by Windows Media Center. This package also provides different types of information related to the Media Center command and control that was used not only for the speech interface development, but also to build the grammar for the recognition engines feeding, described in detail later in section 3.3.

3.1.2 Vista Media Center Add-In

The .Net Add-Ins framework supports functional code that can interact with Media Center using the C# programming language. There are two types of Add-Ins for the Media Center: Background and On-Demand.

Background Add-In

A type of software application hosted by Windows Media Center that automatically starts running in the background soon after Windows Media Center is started, and continues to run until the application closes on its own or is forced to close because Windows Media Center is closing.

On-demand Add-In

A type of Windows Media Center hosted application that starts running only after the user invoke it from the Windows Media Center shell. An on-demand application performs a discrete, synchronous task and then closes when the task is completed.

In this project, we have used a background Add-In that runs at Media Center start-up and remains running in the background. It is through this component, that speech recognition and synthesis is processed.

At start-up, the Add-In starts with the speech recognition grammars creation (to be included in the recognition engines), one for the Portuguese engine and another for the English engine, where all main commands like "music", "pictures", "main menu", "videos", "volume up", "volume down", "next music", etc, and all media information like media names and media tags are included. The grammar creation is described in detail later in section 3.3.

After the grammar creation, both SR engines (ENG + PTG) are initialized, and will be working concurrently.

At runtime, after the grammar is loaded by the recognition engines, this Add-In provides users a speech interface for Media Center command and control. The system is then ready to start receiving speech commands in Portuguese and/or English. When a speech event is received by the application, both engines receive it.

Ambiguity scenarios in the recognition engines selection will not happen in our case, because there are no similar commands in the grammars. For example, in the Portuguese grammar we have "música" and in the English grammar we have defined "music library", for the same command. Such differences in the grammar definition are preventing false positives in the recognition process.

3.1.3 Media Player SDK and Media Database

Available media information is retrieved from the host, by using the Media Player SDK [MPSDK07]. This SDK is public and is one of the components of the Microsoft Windows Vista SDK [WVSDK07].
Using this Media Player SDK, our Media Center Add-In consults the host media database and dynamically builds the SR grammar. In this database we can find all the required media information that is useful to identify each one of media files present in the host machine, like media names and tags. All this media information will be included in the Portuguese and English grammars at add-in start-up, and will be described in detail later in section 3.3.

3.2 The Hardware Description

For this project development we have used several different hardware devices. We used a Clasus [CLASUS07] machine with a 2.2 GHZ CPU, 1 GB ram and 80 GB disc with Windows Vista Ultimate edition installed; keyboard and mouse; a CRT monitor; and finally one headset with noise suppression, ideal for speech recognition.

We also used a Microphone Array, technology described in detail later in section 3.3.2 Microphone Array.

3.3 The Media Center Add-in architecture

In this chapter we present the software architecture in Figure 12. In the following chapters we described the most important components of the system that will be helpful to explain all the software architecture.

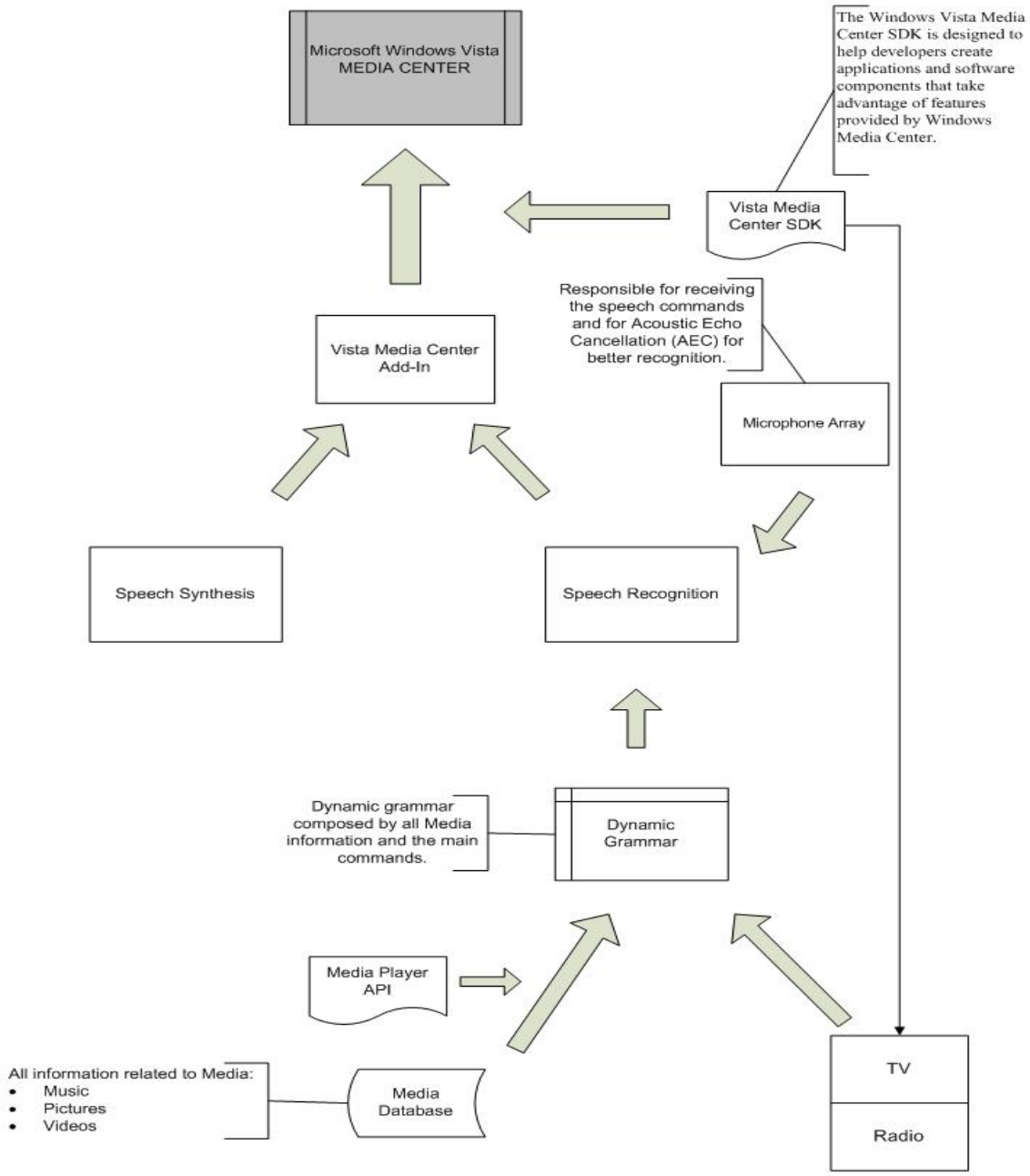


Figure 12 – The Media Center Add-In architecture

3.3.1 Speech Recognition and Speech Synthesis

Our system uses Microsoft desktop Speech Recognition (SR) with Portuguese and English acoustic models, for command and control, to receive and understand user's spoken commands. These two engines will work simultaneously.
The system adopts also Microsoft TTS (Text-to-Speech) technology, for those two languages, for a better interaction between users and the Media Center.
For English synthesis we have used *Microsoft Anna* and for Portuguese synthesis *Madalena* from Nuance.

For all system development, including these speech technologies (recognition and synthesis) we used Microsoft Speech FX library [SFX07] included in .Net Framework 3.0.

3.3.1.1 Speech Corpus and Acoustic Model Used in the European Portuguese Recognition Engine

Our European Portuguese Recognition Engine, recently developed by the MLDC team (Microsoft Portuguese Speech Team) use an acoustic model trained with more or less 95 hours of audio, divided in two major corpus. All the information bellow is all the information that we have for these corpus.

- **Phil 48**

  20 024 waves
  1384 speakers – 10 phrases for each speaker
  16 bits-linear 8 KHz
  7 hours of aligned speech

- **SpeeCon**

  Trained with 87 hours of aligned speech
  266 male speakers and 287 female speakers (Total: 553 speakers)
  Four speech channels (four different microphones placed at different distances) is recorded at 16 kHz with 16 bit quantization

We will now present three tables containing Speaker Accent Regions (Table 6), distribution of adult speakers over regions (Table 7) and distribution of adult speakers over age groups and male and female distribution (Table 8).

Southern
Central
Northern

Table 6 - Speaker Accent Regions for Portuguese (Portugal)

Number	Name of accent/region	#Male speakers	%	#Female speakers	%
1	**Southern**	138	25.0%	152	27.5%
2	**Central**	65	11.8%	65	11.8%
3	**Northern**	63	11.4%	70	12.7%
TOTAL		266	48%	287	52%

Table 7- Distribution of adult speakers over regions

Age groups	Number of male speakers	Number of female speakers	Percentage of total
15-30	127	143	48.8%
31-45	91	102	34.9%
46+	48	42	16.3%
TOTAL	266	287	100%

Table 8 - Distribution of adult speakers over age groups and male and female distribution

3.3.2 Microphone Array

Knowing that the usage scenario for this platform is a noisy scenario (living room), not only related to people conversation but also to the audio output from the system, we have decided to integrate this new Microsoft technology in the system.
This technology (Figure 9) was developed by a Microsoft Research [MSR07] team headed by Ivan Tashev [ITASHEV07].

A microphone array is a set of closely positioned microphones. The latter can achieve better directionality than a single microphone by taking advantage of the fact that an incoming acoustic wave arrives at each of the microphones at a slightly different time.
In real time, the microphone array engine searches for the speaker position and acts as if it points a beam at the current speaker. The higher directivity of the microphone array reduces the amount of captured ambient noises.
With the integration of this technology in Media Center Controlled by Speech project, the system should be able to capture user's speech orders using Acoustic Echo Cancellation (AEC), and Noise Reduction, trough Automatic Gain Control, which are already incorporated in Microsoft Microphone Array technology, primitively supported in Windows Vista, but not yet in Speech Recognition. After this integration, the system should be able to capture user's speech commands for robust speech recognition of a speaker in the presence of ambient noise and/or music.
The ideal scenario is the system playing some music or some video, and the user sending some order. In this case, using Microphone Array, the recognition engine just receives the user order, because all the system noise is automatically cancelled using AEC. The multi-speaker scenario is not supported yet.
The integration of the latter in the ASR is completed (1st phase) but the integration in the Media Center add-in (2nd phase) is still ongoing, but the main issues were now solved.
The MicArray is accessed through a Direct Media Object (DMO) that does all the processing in runtime and basically gives us a byte array buffer in a continuous loop.
The goal was to take the DMO audio output, the byte array buffer, as an input into a SR engine continuously. For that, we first tried to use a custom audio object, as described in the SAPI

whitepapers, but we've found that it would be easier to use a MemoryStream as input to the SetInputAudioStream member function in SpeechRecognitionEngine.

Since the SR engine reads back data from the stream faster than the DMO output processing routine can write the SR engine will soon finish reading all data. The problem was that the SR engine don´t resume reading from the stream once more data is available. It simply stops recognition.
To solve this problem we simply control the speed at which the SR is reading from the stream and keep it from shutting down when it reaches the end of the stream.
Our approach was: we derived the MemoryStream class and overridden the Read and Write methods. Basically, when the SR calls Read and there's no more data at the moment, we temporarily lock it in until there's enough data. When the DMO writes new data to the stream the SR resumes reading.

3.3.3 Dynamic grammar

The ENG and PTG recognition engines, each require their own context free grammar that is feed by the main speech commands for Media Center control, but also all the media information available in the host machine, such as, media name, tags, subject, artist, genre, year, etc. The grammar allow users to speak specific commands, such as: "play rock music", "show 2007 pictures" or even "play 2004 vacations video".

Each one of the recognition engines has their own grammar. The Portuguese grammar has Portuguese entries/commands like "menu principal", "música", "videos", "imagens", "tocar estilo rock', "tocar Sting",etc. On the other hand, the English grammar has English entries/commands like "main menu", "music lybrary", "videos library", "pictures", "play genre rock", "play Sting",etc.

The system should permit users to add new media files at runtime. For this reason both grammars are constructed dynamically.
The relevant media information is gathered at runtime, and can even be updated during the application execution. At start-up, all the system main commands are added to the grammar (e.g. "music library", "TV", "radio", "pictures", "movies", "more programs", "main menu", "volume up", "stop", and so on). After that stage and after retrieving all media information through the Media Player API (described later in section 4.2), more complex commands (defined by us) are added to grammar to control all available media, Still Pictures, Movies and Music. Other commands to control the TV, the Radio and Internet Access, are retrieved through the Media Center API.
All the main commands available in both grammars are listed, in detail in annexes 2 and 3.
The architecture of the developed grammar system module is displayed in Figure 13, where two different API's are used.

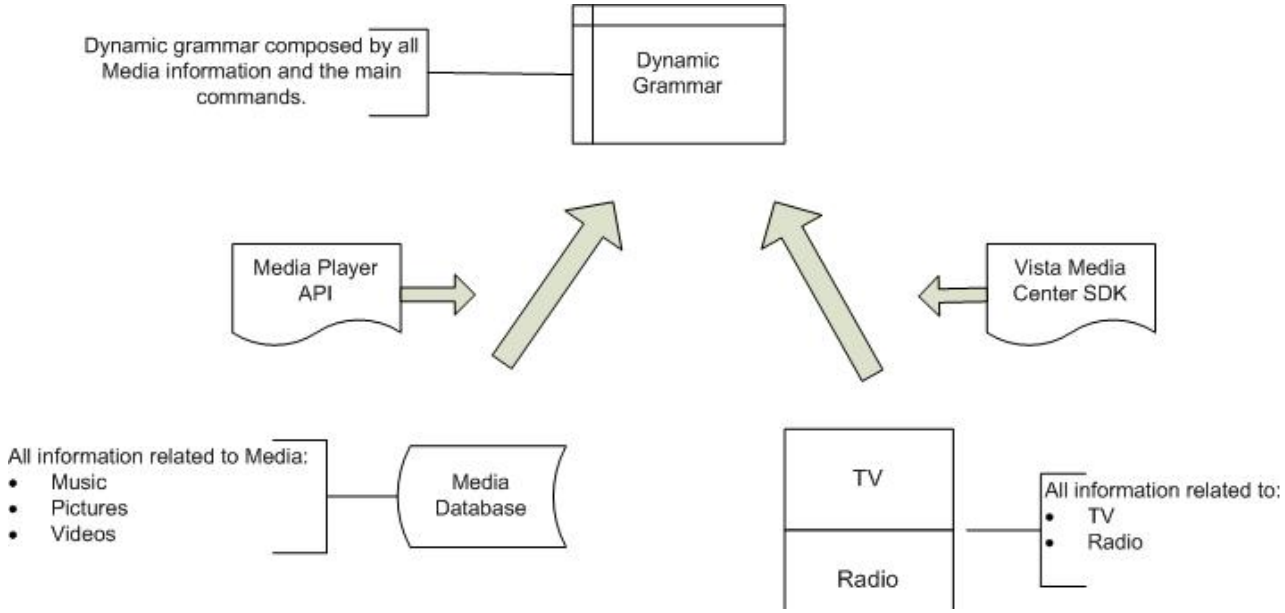


Figure 13 – Dynamic Grammar Process Diagram

3.4 Developed work

Before starting the thesis itself, some other small projects were developed to speech recognition technology and C# programming language initial approach and study.
One of those projects was very useful to solve one of the major thesis problems, the multiple recognition engines problem.

3.4.1 Multiple recognition engines problem

The main difficulty of this problem, inhabited in the impossibility to have two different recognition engines in simultaneous operation, something that was solved with the initialization of the recognition engine with the InProc option.
There are two possibilities to initialize the recognition engine: shared and InProc. In the first option, it is only possible to have an active engine in one specific moment, because the latter is shared by different applications.
In the second option, which was used in this thesis, it was possible to have more than one active engine, as they are exclusive to just one application.
All information related to this application can be found at Annex 7.

3.4.2 EVA – Easy Voice Automation

A new Microsoft technology was also studied during this thesis- the EVA (Easy Voice Automation) system.

One of the biggest advantages of the speech interface for Media Center developed during this thesis was the usage of Speech Macros. These Macros allow users to trigger complex actions using just one word. This concept is also used in EVA.

EVA is a new Microsoft project, using XML language, which attempts to reach a simple way of adding voice commands to any Windows program, without modifying them.
This is possible by running an application in background (voice recognition), with the capability of start programs and interact with them and/or send commands to these running applications.

Before we start using Media Center SDK to develop the speech interface for Media Center, we tried to use EVA to send speech orders to the system, but we quit this choice because it would became more difficult and complex than using the MCE SDK. Using EVA we would also have to use Media Center SDK to send specific commands to the system, and for this reason we would have the double of the work than using just the MCE SDK.

A briefly description of this technology is included in Annex 4.

3.5 Conclusion

In this chapter we have presented the project developed details, including a complete list of software, libraries and SDK's used for the application development. In this chapter is also described all the software architecture and the description of each one of the principal modules.
Finally, is also described the Microphone array technology as well as the integration in the SR engine.

4. Usability Evaluation

4.1 Introduction

Usability evaluation is a very important tool to determine the usability of the system, and can be conducted at many stages during and after the design and development process.
Our main goal with these tests and in this stage, was to compare the two different interfaces available to control the system, the usual remote-control and the developed (speech interface).
Comparing these two interfaces will determine their advantages and disadvantages, strengths and weaknesses, as well as how the different experienced end-users adapt to them, especially to the speech interface.

The main goals to achieve with the execution of the usability tests were:

- Validation of the level of acceptance of speech recognition technologies in European Portuguese on Media Center.
- Obtain fast turnaround feedback from local Media Center users at the Microsoft Portugal subsidiary and implement it in the product.
- Detect what are the "local" needs in this area.
- Allow the users to really have a real experience with the system and hear their positive and negative feedback.

4.2 The interfaces: Remote Control vs Speech

Media Center platform has a simple user interface that is totally controlled by a traditional remote control (Figure 5).
The system can now also be controlled by a speech interface (previously described in chapter 2), which was developed during this thesis, regarding MC full control using only the users pronounced speech.

4.2.1 Usability evaluation methodology

The usability testing experiment was design to assess the advantages of the speech interface comparing with the traditional remote control interface.
To accomplish this goal we have decided to elaborate a list of tasks (Table 9), to be executed by the users. With this approach we tried to evaluate the way users interact with the system and how they perform some simple tasks like playing some music or see some pictures, with the speech interface (Figure 14) and with the remote control.

	Tasks
1	Go to the music menu and play an album
2	Change music, raise volume , stop music
3	Go back to the music menu and play an artist
4	Stop music and go to the main menu
5	Go to the videos menu and play a video
6	Stop the video and go to the main menu
7	Play some music and while music is playing see the photos of Afonso

Table 9 – List of task to the usability evaluation

Figure 14 - Speech interface

Before performing their tasks, users have received 5 minute training sessions for each different interface, where the main commands and basic menu navigation were explained.
Before and during the usability tests, users could consult the user-guide (annexes 2 and 3), or use the system help (Figure 15).

Figure 15 – System help

Before execute the tasks, we let users freely test the system, not only for a system adaptation but also to collect other speech commands synonymous to include in following versions.

4.2.2 Usability experiment

The usability experiment was run on 35 volunteer users (all from our Microsoft subsidiary).
We have tried to get a large number of users from different areas and ages (Chart 1), not only from informatics but also from financials, human relations, marketing and so on, because our goal was to get feedback not only from the Media Center and speech interfaces common users, but also from users that were even unaware from the existence of these technologies (Chart 1 and Chart 2).

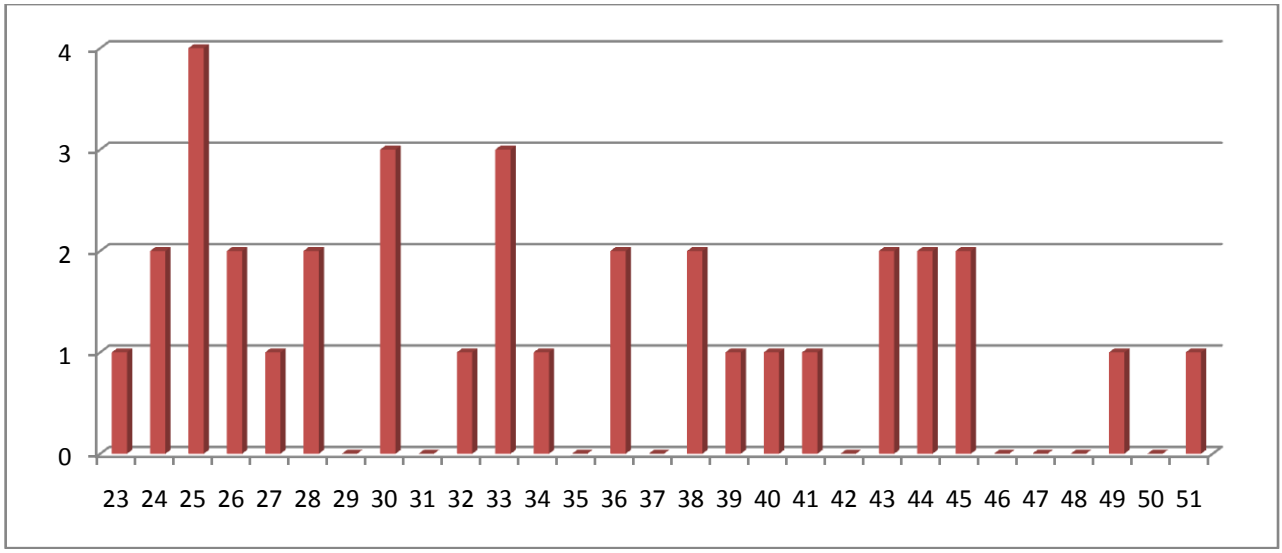


Chart 1 – Subjects age

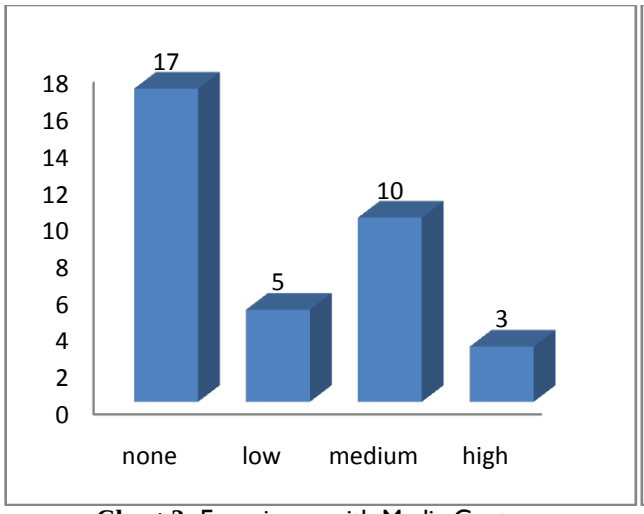


Chart 2- Experience with Media Center

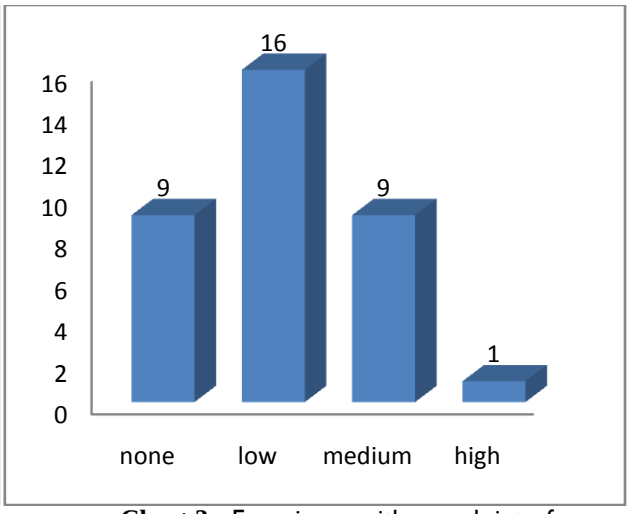


Chart 3 - Experience with speech interfaces

For this reason, we first retrieve the user's expectations and experience regarding speech interfaces through a preliminary subject questionnaire (Table 10).

NAME:
Age:
Genre:
What is your experience with speech interfaces?
What is your experience with Windows Vista Media Center?
Which Windows version do you use at home?
Which Windows version do you use at work?
In your opinion what should be the main advantages introduced by a speech interface for this kind of systems?

Table 10 – Preliminary Subject Questionnaire

The next step was to explain the Media Center platform to users and let them freely test the system with the remote control or with speech. This step was made alternated, where one user test the system with the remote and the next user with the speech, and so on. This way, we could balance both tests, and have a similar number of subjects testing both interfaces.

The tests were also alternated between the two languages supported, because we also pretended to assess the system usability using the subject's mother language and comparing the results with the English version. The first subject has executed his usability test using the Portuguese add-in, the second subject with the English add-in, and so on. This way we could also balance both tests, and have a similar number of non-sequential subjects testing both interfaces. The subjects number 1,3,5,7,9,… tested the system in Portuguese and the subjects number 2,4,6,8,… tested the system in English.

After the system adaptation, we gave to users a list of tasks to be executed with the remote control and with speech. We have previously calculated an optimal number of speech commands for each task, as showed in Table 11.

	Tasks	Expected number of speech commands
1	Go to the music menu and play an album	3
2	Change music, raise volume , stop music	3
3	Go back to the music menu and play an artist	2
4	Stop music and go to the main menu	2
5	Go to the videos menu and play a video	2
6	Stop the video and go to the main menu	2
7	Play some music and while music is playing see the photos of Afonso	4

Table 11 – Tasks executed and expected number of speech commands

This step was also been made alternated. Subjects number 1, 3, 5, 7,… had tested the system using speech in a first phase and using the remote control in a second phase, and the rest of the subjects had tested the system using the speech interface first and then using the remote control. This way we avoid users to get experienced with the system using the first tested interface and then it would be easier to complete the tasks using the second interface.
While users here executing these tasks (Figure 16), we here collecting all information regarding problems and issues found, new speech commands synonymous and collecting the time spent to perform each task.

Figure 16 – Subjects performing their usability tests

Finally, the last step was to retrieve user's feedback, not only for the new speech interface but also comparing it to the remote control. For these tasks, users have responded to another questionnaire (Table 12).

Name:
What is your general opinion about this product? (1 - too bad; 10 - superb)
Media Center Controlled by Speech. In your opinion how will common users adapt to this system?
And your sons? (If you have any)
And aged persons, like your parents (p.e.)?
Did the system understand you well?
Did you used 'help' to receive system support?
If you asked for help, did the system help you correctly?
If the 'help' was not good enough, what do you think that could be better?
Which version did you prefer, speech recognition or remote control?
System speed? (1 - too slow; 10 - too faster)
And system reliability? (1 - too bad; 10 - superb)
Changes and innovations that you would like to see in the system?
What bugs have you found in the system?
Free commentaries

Table 12 – Final Questionnaire

4.2.3 Evaluation results and analysis

All the 35 test subjects have successfully completed their tasks. Some of them were quite faster than others, and some have experienced some difficulties, that were solved after some attempts or after using the system help.

During the experiment, we gather all problems and issues found in both interfaces, which were a few, and a large number of new speech commands synonymous, about 150.

Finally, we have made an analysis on the time users spent to perform the experiences, and an interpretation of all answers given by them for the two questionnaires (Table 10 and Table 12).

4.2.4 Problems in the system

The main problems found during the experiences were related to the lack of a large number of speech command synonymous available for the speech interface. For example, to raise volume, users used "volume up", "louder", "more volume", "up", etc.

Another problem occurred during the experiences was related to the recognition engines itself. Just one subject had some problems for the recognition engines to understand him.

4.2.5 Subject comments

Before, after and during the usability tests, all subject comments were register for later analysis. Some of those comments were very interest and provided a valuable feedback, not only for the speech interface itself but also for new innovations to be included in the system.
The most relevant comments were:

Changes and innovations that you would like to see in the system?

- "Media center should be more powerful";
- "Common English terms in the Portuguese version";
- "Menu delays destroy user experience - media center speed is not up to par with the speech interface";
- "Support a large number of synonymous for the same speech command";

What bugs have you found in the system?

- "The system fails to understand my speech commands for several times. To realize my tasks it was necessary to speak louder and with a non-natural voice". (The subject was a girl and has a particularly voice (hoarse), and we think that for that reason the recognition engines had some difficulty to understand her).

Free comments

- "These tests would be easier for more experienced users of Media Center. This is not my case";
- "The Portuguese Speech Recognition was a success. The system has accepted all my speech commands correctly. I'm very happy to see this system controlled with Portuguese language";
- "Nice experience";
- "Change the button disposal in the remote control."
- "Wow! I hate remotes";
- "Good!"

Here are some other comments from important Microsoft Portugal employees.

- **(Information Technology Support)**: "The Speech Reco in Portuguese was a success! The system recognized ALL my commands quickly and effectively. I'm very happy with this comfortable solution."

- **(Marketing Program Manager)**: "Sometimes Speech Reco is quicker but in other areas the remote control is the quickest IF the user knows well how to use it, as its design is not the typical European one (for instance, the directional keys are usually above the "Play, Pause, etc." group and not below it, and the numbered keys are usually above all the others and not below all them). This different positioning significantly reduces user interaction speed via the remote."

- **(Windows Server Technical Sales Manager)**: "Speech interaction with the PC is much more intuitive than the remote. I prefer doing it in the Portuguese version because the speech and recognition of our native language is much easier and quicker than having the system recognizing the English pronounced by Portuguese users."

- **(ex- Windows Server Program Manager)**: "If I have to interact with the remote I still prefer the EN system but if I have to do it my "voice" (speech), I prefer my native language which is much easier, handier and quicker."

- **(Mobile Business Group Lead)**: "Indeed very interesting and utile experience!! The interaction via Speech and in PTG is done in a more relaxed scenario, at the sofa, with friends, which makes it very attractive. I hate remotes as having to go through complex menu navigation sometimes frustrates. This is quicker and very practical!!"

- **(Windows Business Group Lead)**: "Great initiative to collect end user experience! No doubt this Speech solution is a great add value to the Portuguese market! Major OEMs would be glad to try it out and see the direction we're taking. They'd surely invest on more robust MC solutions to sell in the market. We should seek this new level of alliance."

- **(Business and Marketing Officer)**: "Very comfortable indeed! We should make it available ASAP! When is it on the market?"

- **(Microsoft General Manager)**: "The next step in human/PC interaction: intuitive, natural and very comfortable. Let's keep it up"

4.2.6 Time analysis

While each subject was executing their tasks, we capture the time, using a chronometer, required to complete each one of them, using the speech interface and the remote control.
As we can see in the next chart (Chart 4), time spent to perform the given tasks using the speech interface was quite faster than using the remote control. This difference was bigger in the most complex tasks and similar in the simple tasks.

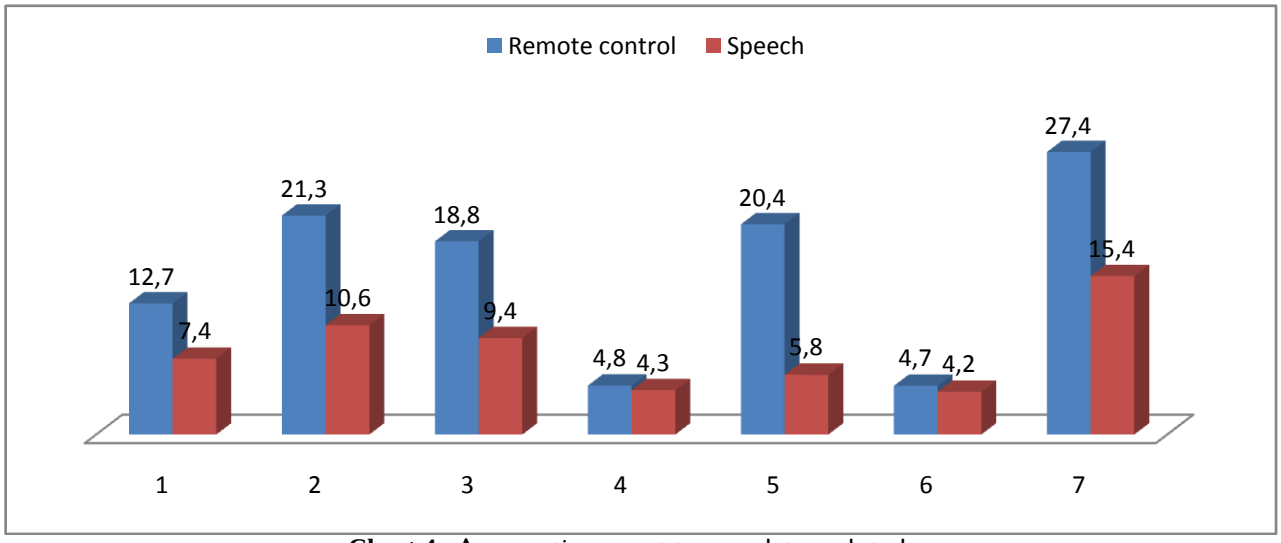


Chart 4– Average time spent to complete each task

After analyzing these results, we concluded that for the simple tasks, there is no relevant time spent difference between the two different interfaces, but for more complex task, the speech interface became more efficient faster and easy to use, as we can see by the average time spent in tasks 2, 3, 5 and 7.
For example, for performing task 5 (Go to the videos menu and play a video) subjects spent almost 4 times more time using the speech interface than using the remote control.
We can also see that, in average, using the speech interface is 42% faster than using the remote control.
The standard deviation for the remote control was 8,65 seconds and for the speech interface was 4 seconds.

4.2.7 Questionnaire and observed results

After the usability tests, we asked subjects to complete their participation by respond to a final questionnaire (Table 12), to collect the maximum feedback with their experience using the speech interface in the Media Center platform and compare this interface to the familiar remote control. The results were very satisfying, as we can see in the following charts.

In Chart 5, we can see that the subject's general opinion about the product with speech interface was positive. Most users (13) rated the product with the highest mark.

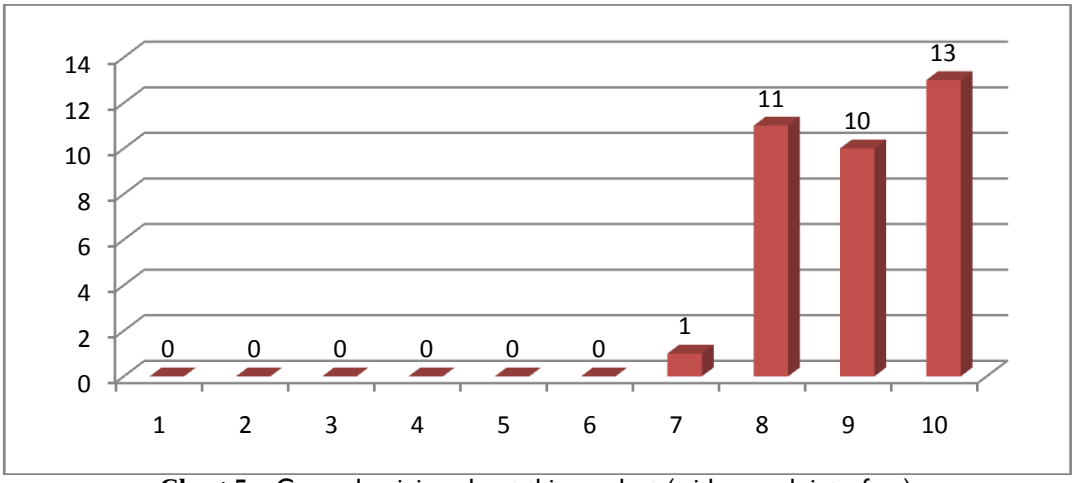


Chart 5 – General opinion about this product (with speech interface)

In the following chart (Chart 6), we can see that most subjects (57%) think that users will easily adapt to the speech interface. These values were related to common users. On the other hand, when we asked subjects the same question but now related to aged users, their think that, comparing with the remote control, the adaptation to the speech interface will be much easier, as we can see in the Chart 7.

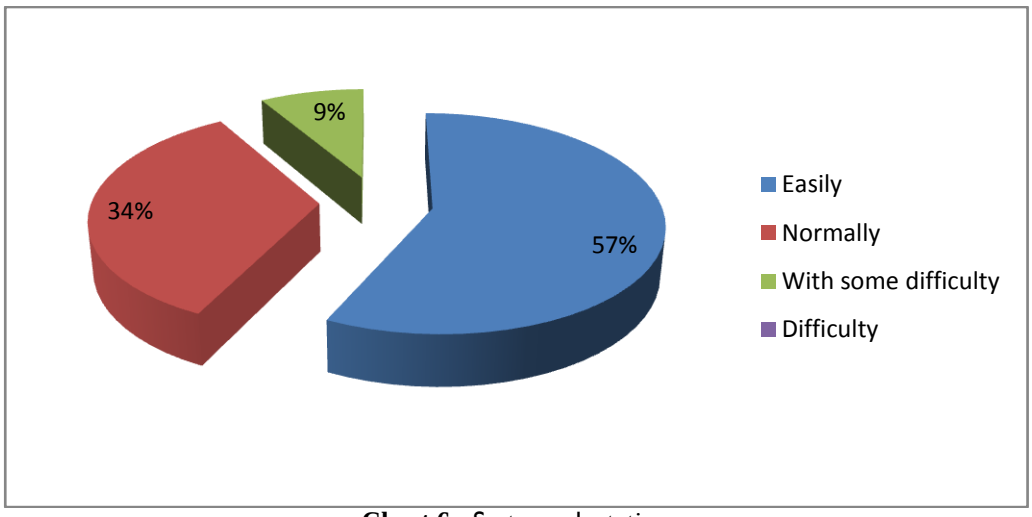


Chart 6 - System adaptation

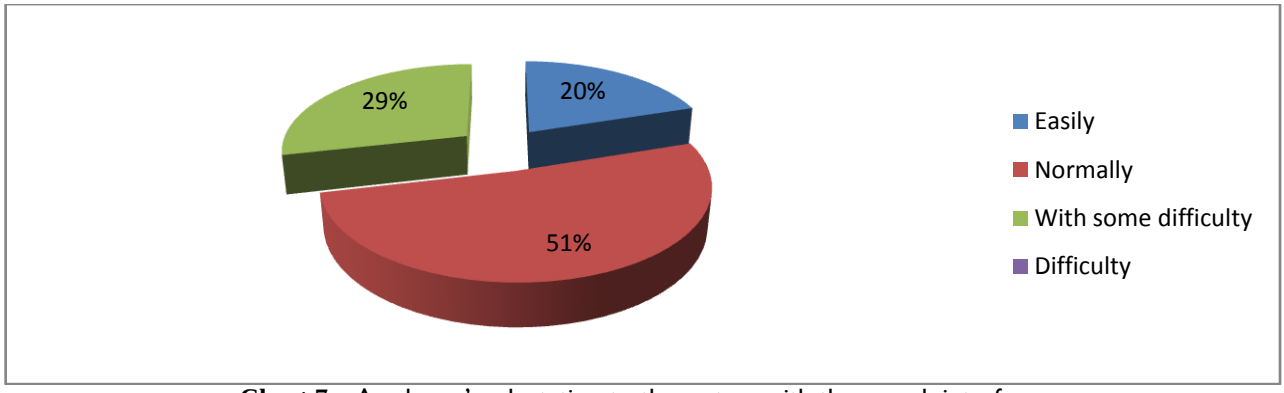


Chart 7 – Aged user's adaptation to the system with the speech interface

Finally, when we asked subjects their preferences between the traditional remote control and the new speech interface, the results were amazing. The graphics results bellow (Chart 8) clearly show that most users (93%) preferred the using speech interface rather than the traditional remote control, even if in the beginning they were reluctant in using speech commands. A small number of them (7%) think that combining both interfaces will be useful. None of the users preferred only the remote control. These results had exceeded our better expectations.

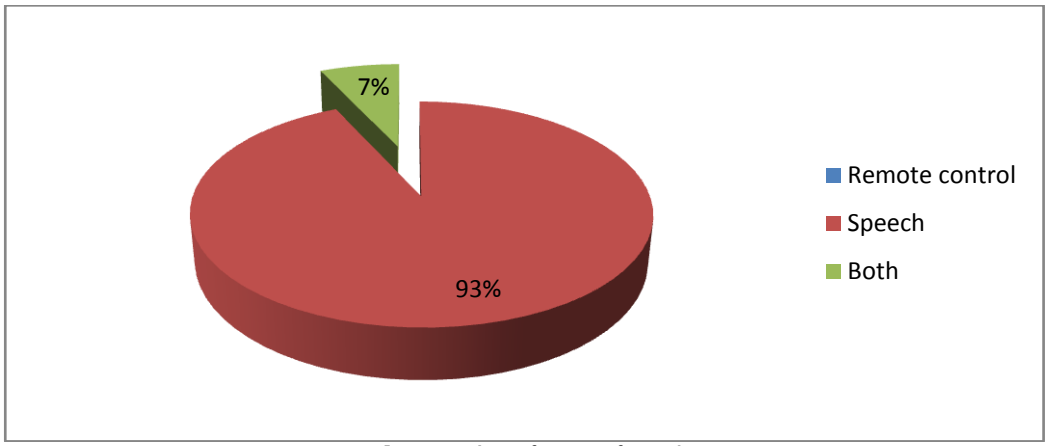


Chart 8 - Interface preferred

4.3 Hoolie VS Speech Macros (Media Center Add-In)

4.3.1 Other usability evaluation

With this new usability test we aim at performing a preliminary comparison between two existent different approaches used to control Windows Vista Media Center, via a Speech interface. The first one is the speech interface developed in this thesis, a specific speech interface for Media Center, that uses a "Speech Macros" approach were a given speech command might correspond to a series of low-level User Interface commands.
The other one is by using Vista Speech Recognition (code-name *Hoolie*). Windows Speech Recognition is a new feature in Windows Vista, built using the latest Microsoft speech technologies. Windows Vista Speech Recognition provides excellent recognition accuracy that improves with each use as it adapts to your speaking style and vocabulary. Speech Recognition is available in several languages and is prepared to control several different applications at the same time, including the Operating System itself.

4.3.2 Usability evaluation

This evaluation was run on 5 users and we have defined the same 7 simple tasks (Table 11), used in Media Center Controlled by Speech usability evaluation tests, to perform with the Media Center and we have registered time spent by subjects to perform each task and counted the minimum Speech commands required for each interface case.

As we can see in Chart 9 and Chart 10, users spent, on average, 8 times more time and used 5 times more speech commands using Hoolie than using MLDC Add-In.

In some situations (tasks 2, 4 and 6), with Hoolie, it was impossible to perform a task, and in that case he don´t show the respective value in the following charts.

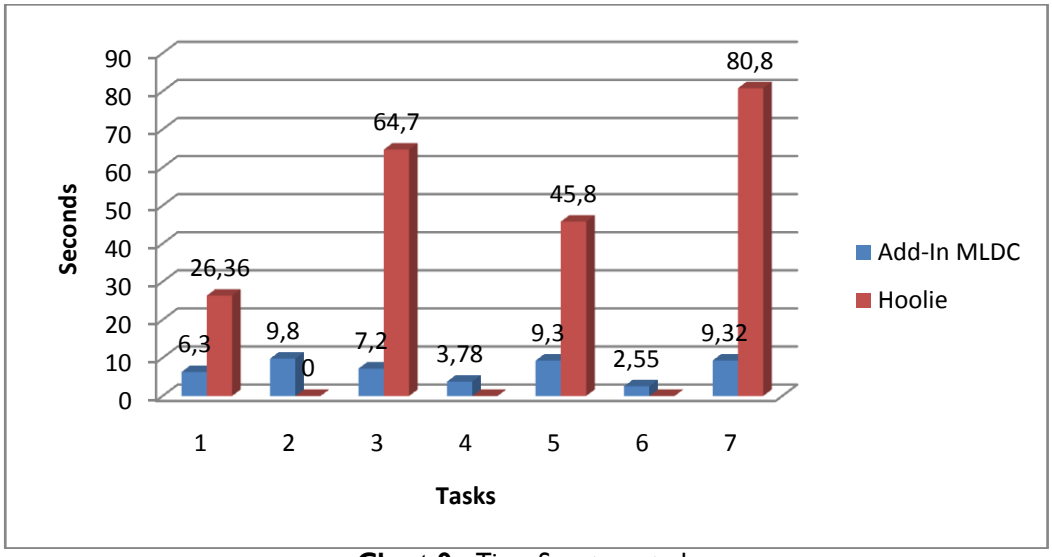


Chart 9 - Time Spent per task

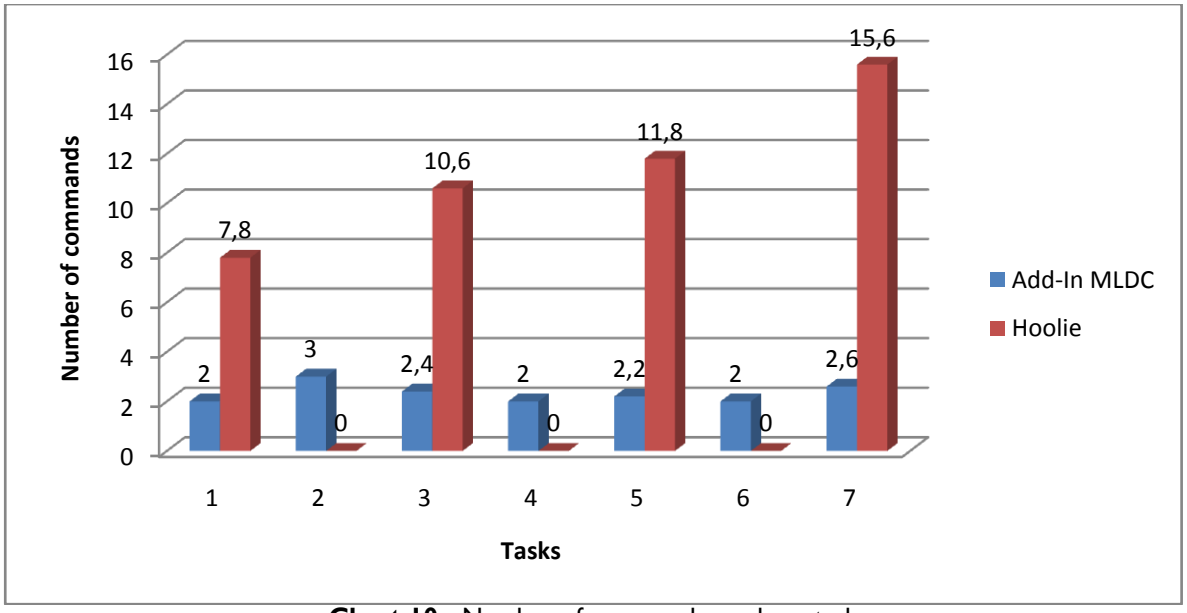


Chart 10 - Number of commands used per task

4.3.3 Users Comments

During the usability tests, all subjects said that there was no advantage in using Hoolie than the speech interface developed during this thesis. Most of them also said that it is almost impossible to control Media Center with Hoolie.

4.3.4 Other issues found

During the usability tests, we have found other important issues that make Media Center almost impossible to control with Hoolie.
Some of them were related to the Media Center interface itself, but the most important were related to the Hoolie performance.

As we know, Hoolie is a generic Windows Vista Speech interface, available for all OS applications and for the OS itself. It is possible to control several different applications at the same time, including the OS itself. For this reason, the grammar used is larger than the expected, because not all the entries in the grammar are specific to Media Center.
When we tried to control Media Center in full screen mode, Hoolie always tried to link speech commands not only with Media Center but also with the OS. Sometimes, during the usability tests, when Hoolie have ambiguous commands, a pop-up message appears behind the Media Center. While Hoolie was expecting a number to proceed to the disambiguation, users continued to send speech commands to the system that were not recognized. This problem causes not only interaction delays but also destroy user experience.

Other problems found:

Main Menu: if the focus is not active in the desired menu, it is impossible to call that menu. For example, as we can see in Figure 17, if the focus is in the TV+Movies, and we pretend to enter in the Music menu, the speech command "Menu" is recognized by Hoolie but it is not executed. To do that, we need to use the primary navigation, using keystrokes: "UP", "DOWN", "LEFT", "RIGHT", "ENTER".

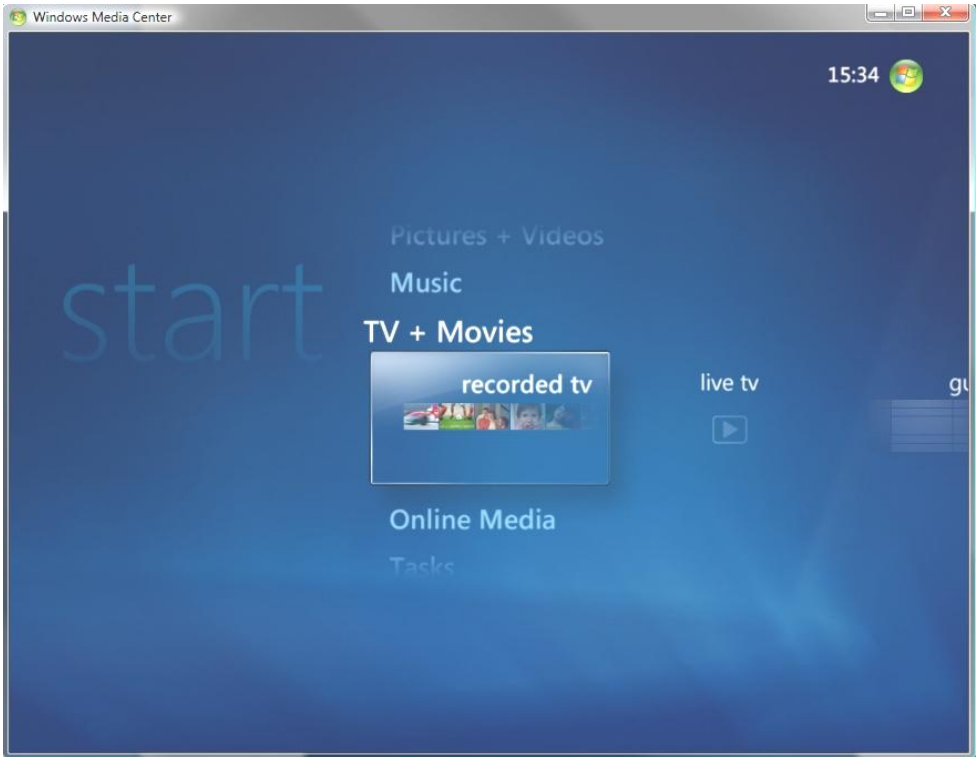


Figure 17 – Media Center Main Menu

Other interesting example is when we try to raise the volume, stop music, play next music or go back to the previous menu. Since those options are not available in the User Interface (see Figure 20), it is not possible to perform those tasks, using Hoolie. When we have tried to say one of those commands, Hoolie will send out an error message like the one in Figure 18. In fact, Hoolie can only perform actions that are available (published) in the Graphical User Interface (Figure 19).

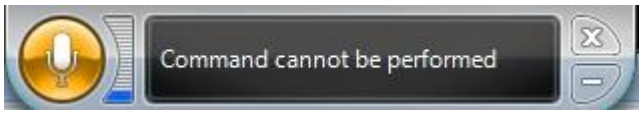


Figure 18 – Command cannot be performed

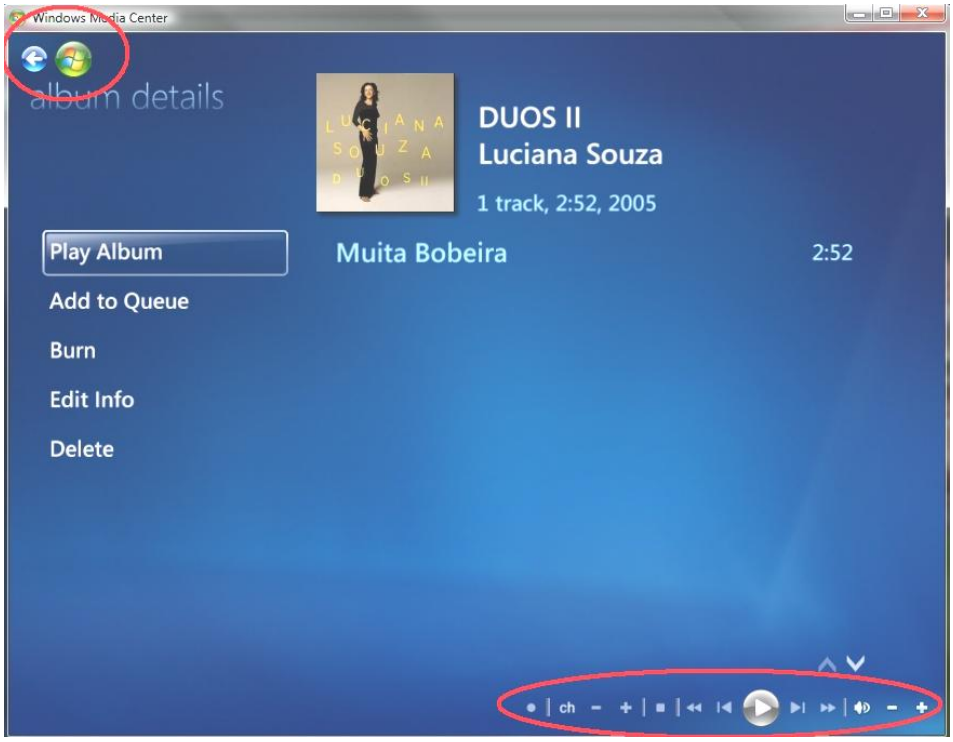


Figure 19 - Options available in the Graphical User Interface

Figure 20 - Options unavailable in the Graphical User Interface

4.4 Conclusion

The results show us that people really enjoyed the experience and reported a great openness for the users to control Media Center with Speech instead of using the remote control, especially because it's more intuitive, supports the direct navigation between different menus (Images, Music, etc.), the speed, the natural interaction by using their mother tongue and a comfortable experience at the sofa.

Some subjects understood another real advantage of the speech interface technology over remote control: they can go directly from an action under a specific area like Videos to a totally different one under another area like Music without having to navigate to the main menu first. This represents a real speed advantage and a more natural interaction paradigm.
The usability tests were a good vehicle to collect feedback (150 new commands only for 3 areas under focus!).

All the remaining information about the usability tests is available in annex 4.

5. Conclusion & Future Work

Conclusion

This thesis describes the architecture and development of prototype of a new speech interface for controlling the Windows Vista Media Center. This interface uses speech recognition and synthesis and provides a simple and natural way to control the system.
All supported input commands, traditionally issued via a remote control, can now be sent to the system via the speech interface, by only pronouncing phrases with no more than 3 words.

Usability tests have shown that the speech interface could be on average 42% faster than the traditional interface. This result opens the possibility for extending in the near future, the usability evaluation experiments to non-experienced users that are sometimes excluded of the usage of advanced audio-visual systems, such as children, elder and visual impaired persons, that may be able to perform not only the simple and obvious commands but also the most complex commands supported by the system, opening this Microsoft platform for all.
We also pretend to perform more usability evaluations, in other Microsoft subsidiaries, regarding a larger system feedback. These evaluations will be performed using the English version.
In the future we also pretend to support other western union languages, in collaboration with the local Microsoft subsidiaries in this region.

Future Work

During the execution of this project we found some problems regarding Media Center TV and Radio support. This problem was related to the Media Center platform itself, more specifically, to the TV tuner cards support. We pretend to solve this problem, as soon as possible, after we get some TV tuner cards certified for Windows Vista Media Center.

To improve user experience, we are also upgrading the media support. It should be possible for users to view their photos, videos and music using other kind of media identifiers, like tags.

As another line of future work, we are enhancing the living room scenario by coping with Acoustic Echo Cancellation, and Noise Reduction, trough Automatic Gain Control, which are already incorporated in Microsoft Microphone Array technology (which include a set of closely positioned microphones)], supported in Windows Vista, but not yet in Speech Recognition.
With the integration of this technology in the Media Center Controlled by Speech, the system should be able to capture user's speech commands with Acoustic Echo Cancellation (AEC) for robust speech recognition of a speaker in the presence of ambient noise and/or music.
The Microphone Array integration is being made in Mono, in this first phase. Later, we will collaborate with Microsoft Research to Stereo integration.

We are also envisaging the enhancement of our PTG acoustic model (regarding better), probably extending it for the case of the support of English words spoken by Portuguese native speakers.

Finally, in collaboration with the local Windows Product Group, we have plans to release a public beta version of this product before the end of the year. This will be a remarkable point for our project.

6. References

[Sharp05] *Sharp, John, "Microsoft Visual C# 2005 – Step by Step, 2005 Edition"*

[Archer05] *Archer, Tom, Whitechapel, Andrew, "Inside C# - Second Edition, Microsoft", 2005*

[Templeman03] Templeman, Julian, Olsen, Andy, "Microsoft Visual C++ .NET – Step by Step, Version 2003"

[SPEECHAPP] Oberteuffer, John A., "Commercial applications of speech interface technology: An industry at the threshold", Human-Machine Communication by Voice colloquium, held by National Academy of Sciences, February 8-9, 1993.

[VOICECOMMAND] Voice Command 1.6, Microsoft, website URL: http://www.microsoft.com/windowsmobile/voicecommand/default.mspx, visited in 26-03-2007.

[WMOBILE] Windows Mobile 5.0, Microsoft, website URL: http://www.microsoft.com/windowsmobile/default.mspx, visited in 26-03-2007.

[SRAPI] Exploring New Speech Recognition And Synthesis APIs In Windows Vista, Microsoft, website URL: http://msdn.microsoft.com/msdnmag/issues/06/01/speechinWindowsVista/, visited in 26-03-2007.

[DEEJAY] The "Dee Jay" system, by Martin L. Shoemaker, website URL: http://tabletumlnews.powerblogs.com/posts/1174189935.shtml, visited in 26-03-2007.

[WMP11] Windows Media Player 11, Microsoft, website URL: http://www.microsoft.com/windows/windowsmedia/player/windowsvista/default.aspx, visited in 26-03-2007.

[WVMC] Windows Vista Media Center, Microsoft, website URL: http://www.microsoft.com/windows/products/windowsvista/features/details/mediacenter.mspx, visited in 26-03-2007.

[NUANCE] Nuance TTS, Nuance, website URL: http://www.nuance.com/, visited in 26-03-2007.

[IVANTASHEV] Ivan Tashev, Daniel Allred. "Reverberation reduction for improved speech recognition", Proceedings of Hands-Free Communication and Microphone Arrays, Piscataway, NJ, USA, March 2005.

[MCE06]: http://www.microsoft.com/windows/products/windowsvista/features/details/mediacenter.mspx, visited at 15.11.2006

[MLDC06]: http://www.microsoft.com/portugal/mldc/default.mspx, visited at 10.11.2006

[MSDN06]: http://msdn2.microsoft.com/en-us/default.aspx, visited at 1.10.2006

[SR06]: http://en.wikipedia.org/wiki/Speech_recognition, visited at 25.9.2006

[SFX07]: http://www.microsoft.com/speech/default.mspx, visited at 25.9.2006

[SS06]: http://en.wikipedia.org/wiki/Speech_synthesis, visited at 25.9.2006

[MicArray06] http://micarray/default.aspx visited at 12.2006

[VMC07]: Windows Vista Media, http://www.microsoft.com/windows/products/windowsvista/features/details/mediacenter.mspx, visited at 11.2006

[MCSDK07]: Windows Vista Media Center SDK, http://www.microsoft.com/downloads/details.aspx?familyid=a43ea0b7-b85f-4612-aa08-3bf128c5873e&displaylang=en, visited at 11.2006

[MPSDK07] : Media Player 11 SDK, http://msdn2.microsoft.com/en-us/library/bb262657.aspx, visited at 11.07.2006

[WVSDK07]: Windows Vista SDK, http://www.microsoft.com/downloads/details.aspx?familyid=C2B1E300-F358-4523-B479-F53D234CDCCF&displaylang=en, visited at 11.07.2006

[CLASUS07]: Clasus, http://www.clasus.pt/, visited at 11.07.2006

[ITASHEV07]: Ivan Tashev, http://research.microsoft.com/users/ivantash/, visited at 11.07.2006

[MLDC]: MLDC, http://www.microsoft.com/portugal/mldc/default.mspx, visited at 11.07.2006

7. Annexes

Annex 1 – Work plan
Annex 2 – Media Center Controlled by Speech European Portuguese User Guide
Annex 3 – Media Center Controlled by Speech American English User Guide
Annex 4 - EVA
Annex 5 – Paper submitted to Interspeech 2007
Annex 6 – Developed Work

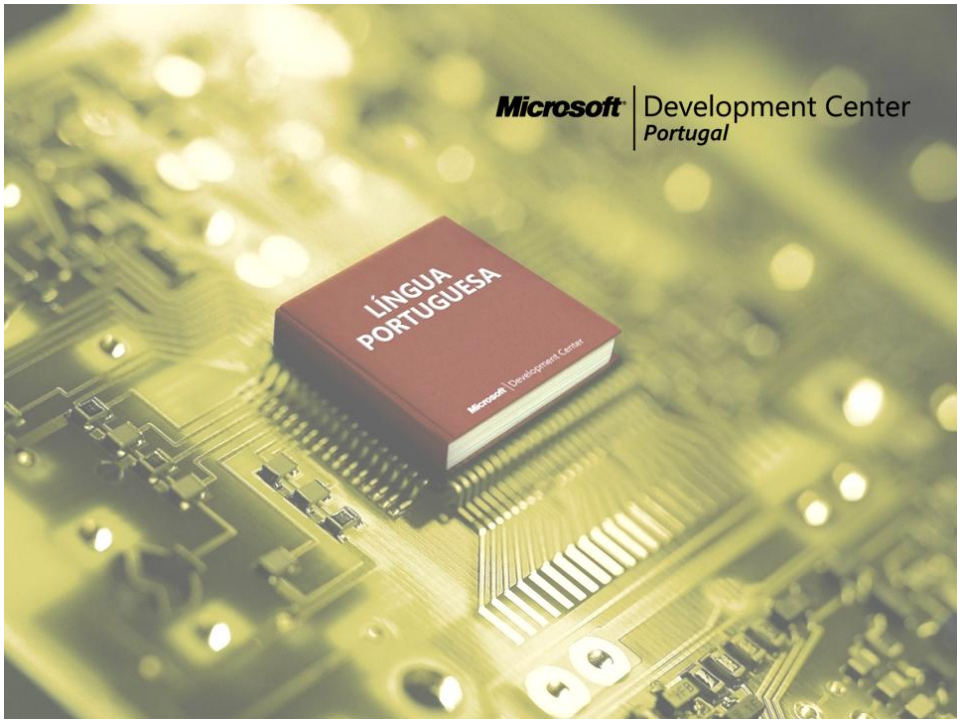
Annexes

Annex 1
work plan

T0 – Início em Setembro de 2006
T0 – T1 (Setembro 06): Análise do problema e estudo da arquitectura e plataformas de desenvolvimento disponíveis para o MS Media Center.
T1 – T2 (Outubro 06): Estudo da SAPI com pacotes de Linguagem Português e Inglês. Realização de um pequeno programa de teste de gramáticas simples nas duas Línguas.
T2 – T3 (Novembro 06): Estudo do problema da detecção automática de Língua falada (PT e UK English), incluindo uma análise do estado da arte e especificação preliminar da solução.
Meta M1 - 30 de Novembro 06: Compreensão da arquitectura e plataformas de desenvolvimento do MS Media Center. Compreensão da SAPI com teste da mesma com gramáticas em PT e UK English. Análise e compreensão do problema da detecção automática de Língua falada (PT e UK English). Entrega do relatório preliminar.
T3 – T4 (Dezembro 06): Especificação pormenorizada do sistema, incluindo o cancelamento da fonte de áudio e a detecção automática da língua falada
T5 – T6 (Janeiro 07): Desenho da solução.
Meta M2 - 30 de Janeiro 07: Especificação e desenho da solução prontas.
T6 – T7 (Fevereiro 07): Desenvolvimento e testes unitário do controlo de fala nas duas línguas.
T7 – T8 (Março 07): Desenvolvimento e testes unitários da detecção automática de Língua falada.
T8 – T9 (Abril 07): Avaliação da usabilidade do demonstrador e produção de 1 artigo curto a ser enviado para conferência Nacional.
Meta M3 - 10 de Abril 07: Desenvolvimento, teste unitário e avaliação de usabilidade do sistema prontos. Entrega do relatório final para revisão.
T9 – T10 (Maio 07): Finalização do relatório final e finalização de 1 artigo curto a ser enviado para conferência Nacional.
Meta M4 - 7 de Maio 07: Entrega do relatório final.
Valoriza-se a produção de um 1 artigo curto a ser enviado para conferência Nacional.
Junho/Julho de 2007. Discussão e avaliação do projecto.

Annex 2
User-Guide in European Portuguese

MEDIA CENTER CONTROLADO POR FALA



MANUAL DE UTILIZADOR

Microsoft Language Development Center, Portugal

COMANDOS EM PORTUGUÊS EUROPEU

Este documento visa ajudar o utilizador da plataforma Media Center controlado por fala em Português Europeu.

Em seguida vão ser listados todos os comandos gerais, isto é, que estão disponíveis a qualquer altura. Numa segunda fase, são enumerados outros comandos disponíveis pra cada menu.

- **COMANDOS GERAIS (SEMPRE DISPONÍVEIS)**
  - **AJUDA**
    - ajuda
    - ajuda-me
    - o que posso dizer
    - o que é que eu posso dizer
    - que posso dizer
    - comandos disponíveis
    - o que é que queres que eu diga
  - **RECONHECIMENTO**
    - Parar reconhecimento
    - Começar reconhecimento
  - **NAVEGAÇÃO ENTRE MENÚS**
    - Voltar (volta para o menu imediatamente anterior)
    - Vídeos
    - música
    - menu principal
    - mais programas
    - programas
  - **MEDIA A PASSAR/TOCAR**
    - Parar
    - Stop
    - Pausa
    - Continuar
    - Reproduzir
  - **IMAGENS**
    - <TÍTULO_FOTO>
    - Mostrar Fotos <das/da/dos/do/de> <TÍTULO_FOTO>
    - Mostrar Fotografias <das/da/dos/do/de> <TÍTULO_FOTO>
    - Mostrar Imagens <das/da/dos/do/de> <TÍTULO_FOTO>
    - Passar Fotos <das/da/dos/do/de> <TÍTULO_FOTO>
    - Passar Fotografias <das/da/dos/do/de> <TÍTULO_FOTO>
    - Passar Imagens <das/da/dos/do/de> <TÍTULO_FOTO>
    - Fotos <das/da/dos/do/de> <TÍTULO_FOTO>

- Fotografias <das/da/dos/do/de> <TÍTULO_FOTO>
- Imagens <das/da/dos/do/de> <TÍTULO_FOTO>

- **SOM**

  - aumentar volume
  - mais alto
  - muito alto
  - baixar volume
  - mais baixo
  - muito baixo
  - silêncio
  - som

- **MÚSICA**

  - <NOME_ALBUM>, <NOME_MÚSICA>, <NOME_ARTISTA>, <ESTILO>

  - toca estilo <ESTILO>
  - toca música <NOME_MÚSICA>
  - toca album <NOME_ALBUM>
  - toca artista <NOME_ARTISTA>
  - tocar estilo <ESTILO>
  - tocar música <NOME_MÚSICA>
  - tocar album <NOME_ALBUM>
  - tocar artista <NOME_ARTISTA>
  - ouvir estilo <ESTILO>
  - ouvir música <NOME_MÚSICA>
  - ouvir album <NOME_ALBUM>
  - ouvir artista <NOME_ARTISTA>

  - passar estilo <ESTILO>
  - passar música <NOME_MÚSICA>
  - passar album <NOME_ALBUM>
  - passar artista <NOME_ARTISTA>
  - passa estilo <ESTILO>
  - passa música <NOME_MÚSICA>
  - passa album <NOME_ALBUM>
  - passa artista <NOME_ARTISTA>
  - artista <NOME_ARTISTA>
  - estilo <ESTILO>
  - música <NOME_MÚSICA>
  - album <NOME_ALBUM>

- **VÍDEOS**

  - <NOME_VÍDEO>

  - Passar video <das/da/dos/do/de> <NOME_VÍDEO>

- Passar filme <das/da/dos/do/de> <NOME_VÍDEO>
- Ver video <das/da/dos/do/de> <NOME_VÍDEO>
- Ver filme <das/da/dos/do/de> <NOME_VÍDEO>
- Mostrar video <das/da/dos/do/de> <NOME_VÍDEO>
- Mostrar filme <das/da/dos/do/de> <NOME_VÍDEO>
- Video <das/da/dos/do/de> <NOME_VÍDEO>
- Filme <das/da/dos/do/de> <NOME_VÍDEO>

- MENU PRINCIPAL

Comandos Específicos Disponíveis:

Música
Vídeos
Imagens
Mais programas

- MÚSICA

Comandos Específicos Disponíveis:

Reproduzir todas
Álbuns
Intérpretes
Canções

- MÚSICA A TOCAR

Comandos Específicos Disponíveis:

Visualizar (Voltar)
Reproduzir apresentação (Voltar)
Música seguinte
Música anterior
Parar
Tocar

- VÍDEO A CORRER

Comandos Específicos Disponíveis:

Pausa
Continuar
Avançar
Retroceder

- IMAGENS

Comandos Específicos Disponíveis:

VER IMAGENS <NOME>
FOTOGRAFIAS <NOME>
FOTOS <NOME>

- SLIDE-SHOW DE IMAGENS

Comandos Específicos Disponíveis:

Voltar

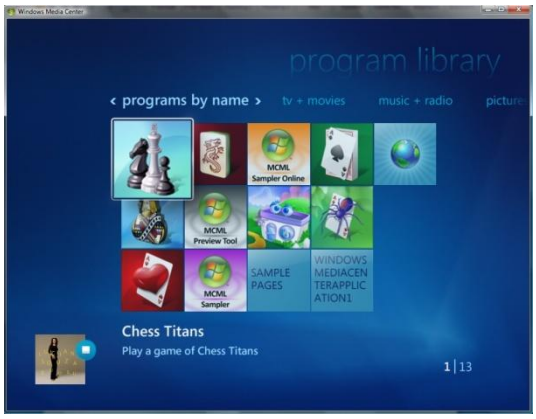
- VÍDEOS

Comandos Específicos Disponíveis:

VER VÍDEO <NOME>
FILME <NOME>

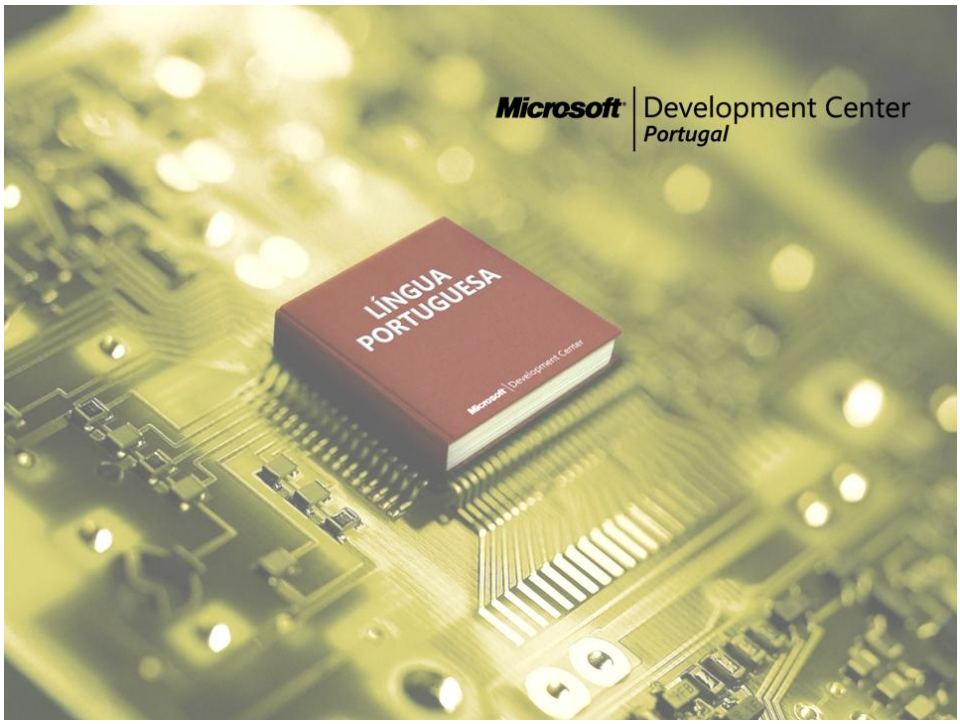
- MAIS PROGRAMAS

Comandos Específicos Disponíveis:



Annex 3
User-Guide in American English

MEDIA CENTER CONTROLLED BY SPEECH



USER GUIDE

Microsoft Language Development Center, Portugal

SPEECH COMMANDS

This document will try to help Media Center users to interact with the system using only their pronounced speech.

We will now list all available commands for each function and menu.

- **General Commands (always available)**

  - **Help**

    - help
    - what can I say
    - help me

  - **Menu Navigation**

    - Back (return to the latest menu)
    - Videos
    - Music
    - Main menu
    - More programs
    - Programs

  - **Pictures**

    - Show Photos <from/of > <FOTOS_NAME>
    - Show Pictures <from/of > <FOTOS_NAME>
    - Watch Photos <from/of > <FOTOS_NAME>
    - Watch Pictures <from/of > <FOTOS_NAME>
    - Watch Imagens <from/of > <FOTOS_NAME>
    - Photos <from/of > <FOTOS_NAME>
    - Pictures <from/of > <FOTOS_NAME>

  - **Sound**

    - Volume up
    - Louder
    - Volume Down
    - Softer
    - Mute
    - Unmute

  - **Music**

    - <ALBUM_NAME>, <MUSIC_NAME>, <ARTIST_NAME>, <GENRE>

    - Play genre <GENRE>
    - Play music <MUSIC_NAME>,
    - Play album <ALBUM_NAME>

- Play artist <ARTIST_NAME>
- listen genre <GENRE>
- listen music <MUSIC_NAME>,
- listen album <ALBUM_NAME>
- listen artist <ARTIST_NAME>
- hear genre <GENRE>
- hear music <MUSIC_NAME>,
- hear album <ALBUM_NAME>
- hear artist <ARTIST_NAME>

- **VIDEOS**

  - <VIDEO_NAME>

  - watch video <from/of> <VIDEO_NAME>
  - Show video <das/da/dos/do/de> <VIDEO_NAME>
  - watch movie <from/of> <VIDEO_NAME>
  - Show movie <das/da/dos/do/de> <VIDEO_NAME>

- MAIN MENU

Available commands:

Music
Vídeos
Pictures
More Programs

- MUSIC

Available commands:

Play All
Albuns
Artists
Songs

- MUSIC PLAYING

Available commands:

Visualize
Play slide show
Previous Music
Next Music
Stop
Resume

- VIDEO PLAYING

Available commands:

STOP
PLAY

- PICTURES

Available commands:

SHOW PICTURES <NAME>
PICTURES <NAME>
PHOTOS <NAME>

- PICTURES SLIDE-SHOW

Available commands:

BACK

- VIDEOS

Available commands:

WATCH MOVIE <NOME>
VIDEO<NOME>

Annex 4 – EVA

This annex is part of a weekly report, where the EVA system was briefly described.

. Study of Speech Macros: EVA (Easy Voice Automation)

. New Microsoft project, using XML language, trying to reach a powerful way of adding voice commands to any Windows program, without modifying them.

This is possibly by run a XML application in background (voice recognition), with the capability of start programs and interact with them and/or send commands to running applications.

. Commands have 3 parts: Conditions (When to listen), Rule generators (What to listen), Executors (What action to perform);

Example:

```
<speechCommands>
 <command>
  <appHasFocus name="calc.exe"/>      → Condition
  <listenFor>divided by</listenFor>   → Rule generator
  <sendKeys>{/}</sendKeys>            → Executor
 </command>
</speechCommands>
```

. It is possibly to have more than one rule generator and executors in the same command (Macro)

. Rule Generators: already implemented:

- Simple phrase (<listenFor>)
- SAPI 5 grammar format
- Numbers (English only, and it's lame)
- Fonts

. Rule Generators: Future:

- Simple static lists
- On screen elements (MSAA)
- Running applications
- Installed applications
- File names

.Executors: already Implemented:

- Display text feedback
- Send keys
- Run command
- Wait for specified time
- Emulate Recognition
- Speak
- Confirm / Alert
- VBScript / JScript

.Executors: Future:

- Insert Text
- Insert Result
- Prompt
- Prompt with HTML form
- Window management
- Call DLL entry point
- Managed code
- Set named state value

Annex 5 – Paper submitted to Interspeech 2007

Windows Vista Media Center Controlled by Speech

Mário Vaz Henriques [1,2], *Pedro Silva* [1], *Carlos Teixeira* [2], *Miguel Sales Dias* [1], Daniela Braga [1]

[1] Microsoft Language Development Center, MLDC, Portugal
[2] Faculdade de Ciências da Universidade de Lisboa, FCUL, Portugal

t-marioh@microsoft.com, i-pedros@microsoft.com, carlos.teixeira@di.fc.ul.pt, miguel.dias@microsoft.com, i-dbraga@microsoft.com

Abstract

This paper describes the requirements and system architecture of a prototype of Windows Vista Media Center controlled by speech. The prototype also implements new approaches of integration of speech recognition of two major Western Europe languages: English (ENG) and Portuguese (PTG). The recognition engines for these two languages work concurrently during application runtime, making it possible the switch, in sequence, between Portuguese and English speech. Speech technology applied to the Media Center gives the users full system control, regarding different media management (TV, music, still and moving pictures, radio, internet access), through a simple, natural and efficient interface, by simply using their speech. Preliminary objective usability testing, show that the speech interface is 40% faster to use than the traditional remote-control type of Human-Computer Interface (HCI). This fact opens the possibility for more usability experiments targeted to the increase of accessibility of the system oriented to children, elder and visual impaired persons.

Index Terms: Media Center, Speech Recognition, Text-to-Speech, Speech interface

Introduction

Speech recognition and synthesis technologies are already present in the daily life of an increased number of people. Due to the potential of speech, as a natural HCI modality, a large number of applications are now adopting it for their main control interface. Commercial, educational and medical [1] applications are now taking benefits of speech, bringing not only a more sophisticated interface for command and control and sometimes dictation, but also broadening its usage for all kinds of users, including children, elder and visual impaired persons. This technology is nowadays helpful for many professional contexts, home entertainment as well as in the small daily tasks that contribute for the general well-being. Applications such as Voice Command [2] for Windows Mobile [3], Windows Vista Speech Recognition [4] for operating system full control, or telephony server-based applications [1], are some examples of products now available using speech technology, that are useful to simplify people's life. However, as for the home "living-room" scenario, whereas a user accesses the home audio-visual devices via natural Human-Computer Interface (HCI) modalities such as speech, there are very few applications that use speech recognition and synthesis, described in the literature. One such project, referred to as "Dee Jay" [5], is a voice-controlled juke box for Windows Vista: you tell it what song you want to hear; the corresponding song entry is found in your library and the Windows Media Player [6] will play it. It is also capable of placing searches with different criteria: by artist or genre, play albums or

collections, and perform a number of other commands. This application only enables listening to music using Media Player. Our line of work, extends this prior knowledge, by supporting fully the "living-room" scenario, were different kind of media such as TV, radio, still and moving pictures, music and internet access are available.

Windows Vista Media Center [7] (or Media Center in short), is a Microsoft platform for home digital entertainment, with an easy-to-use interface appropriate for all family members, controlled by a traditional remote control as a HCI device. The main use case for this platform is the mentioned living room scenario. By introducing speech in this kind of platform, we are envisaging a more natural and simple control of its functionalities. Spoken commands like "Play Jazz", "Watch channel 4" or maybe "Show 2005 vacation pictures" introduce a natural and easy way to interact with the system, giving a secondary role to the remote control. This kind of interaction increases also the accessibility for users groups that are often forgotten.

Our system uses desktop Speech Recognition (SR) with Portuguese and English acoustic models, for command and control, to receive and understand user's spoken commands. The system adopts also TTS (Text-to-Speech) technology, for those two languages, for a better interaction between users and the Media Center. In this paper we describe how we have developed this new interface for this system, including architectural and usability evaluation issues.

This paper is structured as follows:

- In section **0**, user and system requirements are presented.
- In section **0**, we explain the main system architecture, including the defined ideal usage scenario. In this section we also explain how the Context Free Grammar for SR is constructed.
- In section **0**, the development of the system is explained together with the used software and hardware technologies, libraries and SDK´s used.
- In section 5, preliminary usability tests are shown.
- In section 0, conclusions are drawn and lines of future work envisaged.

User and System Requirements

Microsoft Media Center is a platform for all family digital entertainment at home. The main objective of this system is to concentrate in one single technology the capability to explore and view all audio-visual media. This platform has a simple interface, as showed in Figure 1, which is currently managed by a remote control. The large number of menus and options offered by the system increases the complexity for the use of this traditional HCI (remote control, see Figure 2).

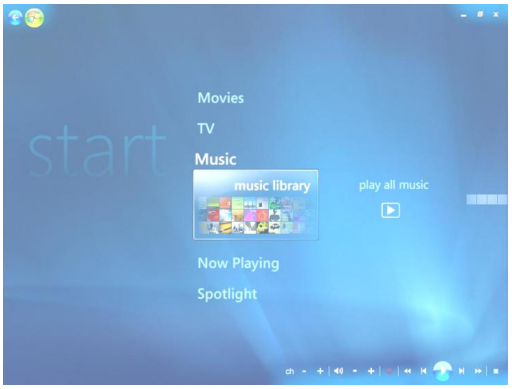


Figure 1 - Windows *Vista Media Center Interface.*

Figure 2 - Media Center Remote Control.

In this section we list user and system requirements needed to develop the speech interface for Windows Vista Media Center.

User Requirements

The main goal of the speech interface for the Media Center is to reduce the user's complexity to interact with the system by adopting a natural HCI. For this reason, the only requirement for system control is the user's speech.

Hardware System Requirements

The Media Center provides support for various kinds of media entertainment through a simple and small hardware component connected to a TV. This may be seen as a powerful PC that must be capable of supporting the required types of digital media for entertainment purposes. For this reason, this hardware must fulfil the following technical requirements:

- Windows Vista Ultimate Operating System
- 1 GHz 32-bit (x86) or 64-bit (x64) processor
- 1 GB of system memory
- 40 GB hard drive with at least 15 GB of available space
- TV tuner card required for TV functionality (compatible remote control optional)
- Support for DirectX 9 graphics with:
  - 128 MB of graphics memory (minimum)
  - 32 bits per pixel
  - DVD-ROM drive
  - Audio Output
  - Internet access (fees may apply)
  -

Software System Requirements

The speech interface for Media Center platform should allow users to have total control of the latter with speech, in command and control mode, in Portuguese, English or both languages in sequence. The user should be able to interact with the system with simple speech commands like "play genre Jazz", "play all Sting music", "show channel 4" or even "show kids pictures".

Usage scenario

The usage scenario for Media Center is the living room, as depicted in Figure 3. This simple audio-visual system is connected to a TV for presentation of audio-visual content and to an audio distribution system (mono, stereo or surround), enabling users to watch and listen to their favorite movie, music, pictures and to access the Internet via a web browser.

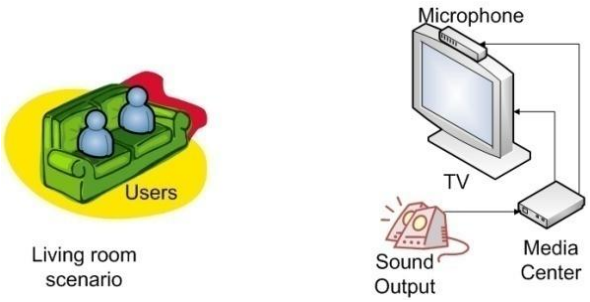


Figure 3: ***Usage scenario.***

Software Architecture

The software architecture, depicted in Figure 4, described its most important components.

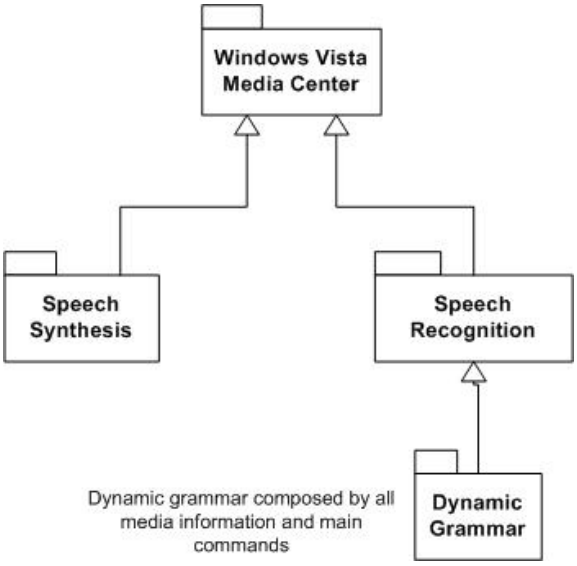


Figure 4: ***Software Architecture.***

Dynamic Grammar

The ENG and PTG recognition engines, each require their own context free grammar that is feed by the main speech commands for Media Center control, but also all the media information available in the host machine, such as, media name, tags, subject, artist, genre, year, etc. The grammar allow users to speak specific commands, such as: "play rock music", "show 2007 pictures" or even "play 2004 vacations video". For these reasons, the grammar is constructed dynamically. The relevant media information is gathered at runtime, and can even be updated during the application execution. At start-up, all the system main commands are added to the grammar (e.g. "music library", "TV", "radio", "pictures", "movies", "more programs", "main menu", "volume up", "stop", and so on). After that stage and after retrieving all media information through the Media Player API (described later in section 0), more complex commands (defined by us) are added to grammar to control all available media, Still Pictures, Movies and Music. Other commands to control the TV, the Radio and Internet Access, are retrieved through the Media Center API.

The architecture of the developed grammar system module is displayed in Figure 5, where two different API's are used.

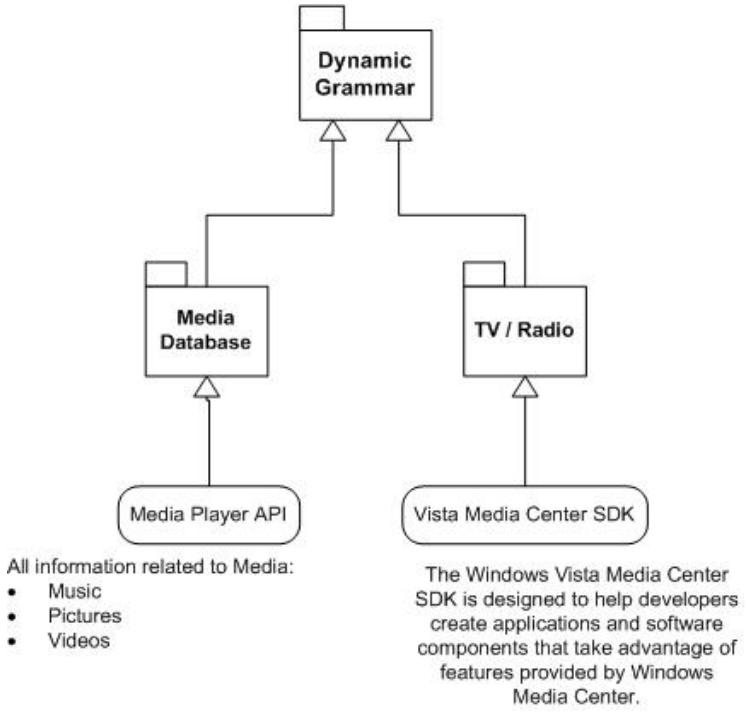


Figure 5: ***Dynamic Grammar.***

System Feedback using Text-To-Speech

The system feedback using TTS provides immediate aid for users, not only when they aren't viewing the presentation feedback, but also for the increase of the perception of the degree of understanding of their speech orders by the system. The system feedback uses available concatenative-based TTS (Text-To-Speech) technologies, for both supported languages. For the English TTS we used the one available in Windows Vista. For the Portuguese TTS case, we used the one licensed from Nuance [8]. Our speech R&D group is currently developing a concatenative-based TTS and we plan in the near future, to incorporate the latter in our Media Center developments.

System Development

For the development of the speech interface for Media Center, a group of technologies were used. The principal component is the Media Center Add-In, where speech recognition and synthesis is made. The complete list of software, libraries and SDK´s used for the application development, is as follows:

- Microsoft Windows Vista Ultimate Operating System
- Microsoft Visual Studio 2005
- Windows Vista Media Center SDK
- Windows Media Player SDK
- Microsoft Speech FX library included on .Net framework 3
- Microsoft Speech API (SAPI)

In the next sub-sections we describe in more detail, the most relevant components used.

Vista Media Center SDK

The Media Center Add-In development requires the Media Center software development kit (SDK). This SDK is public and is designed to help developers to create applications and software components that take advantage of features provided by Windows Media Center. This package also provides different types of information related to the Media Center command and control that was used not only for the speech interface development, but also to build the grammar for the recognition engines feeding, previously described in section 0.

Vista Media Center Add-In

The .Net Add-Ins framework supports functional code that can interact with Media Center using the C# programming language. There are two types of Add-Ins for the Media Center: Background and On-Demand. In our work, we have used a background Add-In that runs at Media Center start-up and remains running in the background. It is through this component, that speech recognition and synthesis is made. At start-up, the Add-In starts with the grammar creation where all main commands and media information are included. At runtime, after the grammar is loaded by the recognition engines, this Add-In provides users with a speech interface for Media Center command and control. The system is then ready to start receiving speech commands in Portuguese and/or English. When a speech event is received by the application, both engines receive it. The decision of who handles the request is made through the confidence returned by both engines. The engine that returns the highest confidence rate is the one who will handle the order. Ambiguity scenarios in the recognition engines selection will not happen in our case, because there are no similar commands in the grammars. For example, in the Portuguese grammar we have "música" and in the English grammar we have defined "music library", for the same command. Such differences in the grammar definition are preventing false positives.

Media Player SDK and Media Database

As described in the previously section 0, available media information is retrieved from the host, by using the Media Player API. This API is public and is included in the Microsoft Windows Vista SDK. Using this API, our Media Center Add-In consults the media database and builds the SR grammar. In this database we can find all the required media information that is useful to identify each one of media files present in the host machine.

Speech Corpus and Acoustic Model Used in the Portuguese Recognition Engine

For the Portuguese recognition engine we have used an acoustic model with WER 7.635% (Word Error Rate). This model was trained with 87 hours of audio (from 157919 wave files, each one with a utterance) and tested with around 10% of that amount (15649 test utterances). The corpus include 266 male speakers and 287 female speakers (in a total of 553 speakers). The corpus was recorded at 16 kHz and with 16 bit quantization, with four speech channels (with four different microphones placed at varying distances).

System Demonstration

A demonstration video (available in [9]) was produced showing the potential of the Media Center speech interface. The system accepts simple commands in both languages (Portuguese/English) in sequence that can perform complex tasks. In fact, all control commands defined by the remote control interface can now be issued via the speech interface, by speaking utterances with no more than 3 words.

Usability Tests

We have conducted some preliminary testing in order to evaluate the usability of the speech interface, by comparing the traditional Media Center remote control and the new speech interface commands in a simple tasks, were we have collected the time duration metric of the tests. The experiment consisted in three simple and different tasks executed in each of the two interfaces, in random order; which were carried out by five subjects. These were colleagues from our R&D lab and Microsoft Portugal: 4 adult male subjects and 1 female with ages between 23 and 35. All of them have been working in the IT field, are daily accustomed with keyboard and mouse and were aware of speech recognition interfaces; 50% of them were not experienced with speech technology. The tasks to complete were defined as follows: playing a specific music, entering the pictures library menu and opening the internet browser. Before performing the task each subject was given a short explanation of how to use both interfaces and he/she was also able to test the system for a few minutes. For each interface across every task, we have collected the duration that each subject spent to complete the task. To accomplish the usability test, we have used the Media Center platform described in this paper, using the remote control for the traditional interaction scenario and a headset for the speech interaction scenario. From Table 1, we can see that the overall duration spent to perform each task was, in average 40% faster using the speech interface than the remote control, which is an encouraging result for our line of research.

users	remote control (seconds) -- speech interface (seconds)		
	playing a specific music	entering pictures menu	opening internet browser
1	4,81 s -- 2,56 s	9,4 s -- 4,59 s	10,98 s -- 2,97 s
2	7,87 s -- 2,43 s	18,04 s -- 11,25 s	10,03 s -- 3,01 s
3	3,78 s -- 2,47 s	11,7 s -- 4,54 s	7,6 s -- 3,6 s
4	10,75 s -- 5,98 s	17,86 s -- 9,22 s	9,9 s -- 4,63 s
5	3,33 s -- 2,79 s	20,11 s -- 4,77 s	20,79 s -- 2,6 s
average time spent	6,108 s -- 3,246 s	15,422 s -- 6,874 s	11,86 s -- 3,362 s
standard deviation	3,14 s -- 1,53 s	4,6 s -- 3,15 s	5,14 s -- 0,79 s

Table 1: ***Usability tests.***

Conclusions and Future Work

This paper describes the architecture and development of prototype of a new speech interface for controlling the Windows Vista Media Center. This interface uses speech recognition and synthesis and provides a simple and natural way to control the system. All supported input commands, traditionally issued via a remote control, can now be sent to the system via the speech interface, by only pronouncing phrases with no more than 3 words. Preliminary usability tests show that the speech interface could be on average 40% faster than the traditional interface. This result opens the possibility for extending in the near future, the usability evaluation experiments to non-experienced users that are sometimes excluded of the usage of advanced audio-visual systems, such as children, elder and visual impaired persons, that may be able to perform not only the simple and obvious commands but also the most complex commands supported by the system, opening this Microsoft platform for all. As another line of future work, we are enhancing the living room scenario by coping with Acoustic Echo Cancellation, and Noise Reduction, trough Automatic Gain Control, which are already incorporated in Microsoft Microphone Array technology (which include a set of closely positioned microphones) [10], supported in Windows Vista, but not yet in Speech Recognition. With the integration of this technology in the Media

Center Controlled by Speech, the system should be able to capture user's speech commands with Acoustic Echo Cancellation (AEC) for robust speech recognition of a speaker in the presence of ambient noise and/or music. We are also envisaging the enhancement of our PTG acoustic model (regarding better), probably extending it for the case of the support of English words spoken by Portuguese native speakers.

Acknowledgements

The authors would like to thank all colleagues that participated in the usability tests and Ivan Tashev from Microsoft Research, for his on-going support in MIC Array technology.

References

[1] Oberteuffer, John A., "Commercial applications of speech interface technology: An industry at the threshold", Human-Machine Communication by Voice colloquium, held by National Academy of Sciences, February 8-9, 1993.
[2] Voice Command 1.6, Microsoft, website URL: http://www.microsoft.com/windowsmobile/voicecommand/default.mspx, visited in 26-03-2007.
[3] Windows Mobile 5.0, Microsoft, website URL: http://www.microsoft.com/windowsmobile/default.mspx, visited in 26-03-2007.
[4] Exploring New Speech Recognition And Synthesis APIs In Windows Vista, Microsoft, website URL: http://msdn.microsoft.com/msdnmag/issues/06/01/speechinWindowsVista/, visited in 26-03-2007.
[5] The "Dee Jay" system, by Martin L. Shoemaker, website URL: http://tabletumlnews.powerblogs.com/posts/1174189935.shtml, visited in 26-03-2007.
[6] Windows Media Player 11, Microsoft, website URL: http://www.microsoft.com/windows/windowsmedia/player/windowsvista/default.aspx, visited in 26-03-2007.
[7] Windows Vista Media Center, Microsoft, website URL: http://www.microsoft.com/windows/products/windowsvista/features/details/mediacenter.mspx, visited in 26-03-2007.
[8] Nuance TTS, Nuance, website URL: http://www.nuance.com/, visited in 26-03-2007.
[9] Demonstration video annexed to this paper.
[10] Ivan Tashev, Daniel Allred. "Reverberation reduction for improved speech recognition", Proceedings of Hands-Free Communication and Microphone Arrays, Piscataway, NJ, USA, March 2005.

Annex 6 – Developed Work

News Reader application

After the voice recognition technology and the development platform with C# programming language initial approach and study, we've started the SAPI (Speech API) analysis.

After this, and to practice the new concepts learned, we've started the development of a new application (*Figure 1 – News Reader*) where users can ask the system to read them sample news in both Portuguese and English, having the capability to receive orders in both languages as well.
This application uses two recognition engines (Portuguese and English) simultaneously, to receive requests from users, and uses TTS to read the news. The system compares the confidence value returned by both engines to distinguish what language the user is speaking. Therefore, the recognition engine that returns the highest confidence value is assumed to be the most appropriate engine.
The application interface is very simple. In the left side, there are some sample categories of news in Portuguese, and in the opposite side other categories in English. To select one of the categories the user just needs to say one of them (e.g.: "Desporto" or "Time"), and the system will then show the full text of the news in the center of the interface. If the user wants the system to read the news, he just needs to say "Ler" in Portuguese or "Read" in English.

To develop this application was necessary to solve one of the major problems previously described, automatic language detection (Portuguese/English).

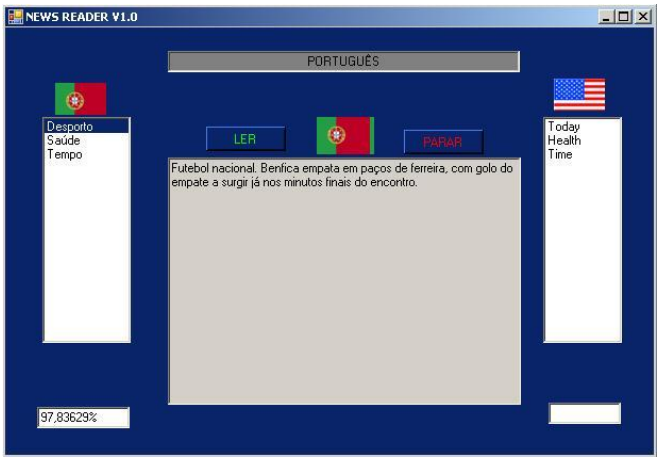


Figure 1 – News Reader

The problem is related with the possibility of a user to be able to use Portuguese or English language to send orders to the Media Center system. This problem was solved after a detailed study of SAPI (Speech API), and testing the application previously developed (*News Reader*). The problem solution is explained next.

Offline Remote Explorer application

Meanwhile, another application has been developed [Figure 2 – Offline Remote Explorer]. This application is a simple tool, similar to Windows Explorer, where the users can retrieve a file system tree directory from a remote site, and save it to file. This allows the user to browse the directory tree in an offline mode, which is the major goal of this application.
To improve the performance multiple threads (managed through a thread pool) were used, allowing a faster retrieval of the remote file system tree. The user can choose the number of threads running in the retrieval process, by changing the number of threads in the top right corner, in the interval of 1 to 99.

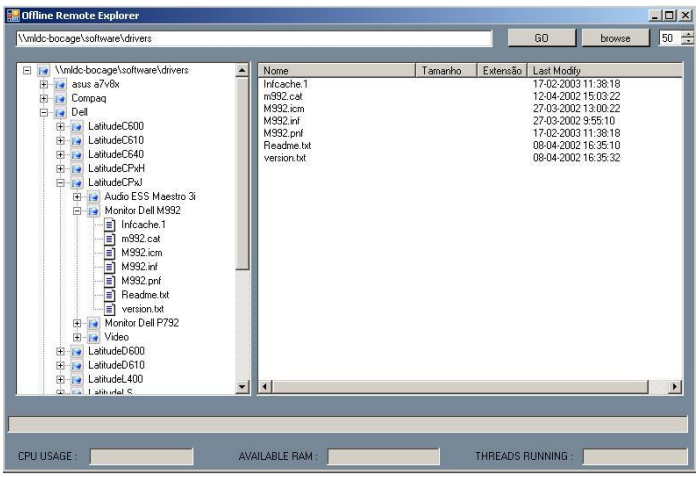


Figure 2 – Offline remote explorer